NORTHWESTERN UNIVERSITY

Essays on the Science of Science and Innovation

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Yian Yin

EVANSTON, ILLINOIS

September 2022

ABSTRACT

Essays on the Science of Science and Innovation

Yian Yin

The increasing availability of large-scale scholarly datasets offers an unprecedented opportunity to understand the fundamental predictability, uncertainty, and dynamics of science and innovation. In this dissertation, I present some of my contributions to the science of science and innovation in three distinct but related settings, through a combination of canonical social science theories, large-scale datasets of science and technology, and mathematical modeling tools. First, I study the quantitative patterns of repeated attempts by NIH investigators, business innovators, and terrorist organizations to build a simple mechanistic model of failure dynamics. The model highlights a novel phase transition that separates failure dynamics into regions of stagnation or progression, predicting that near a tipping point, agents who share similar characteristics and learning strategies may experience fundamentally different outcomes following failures. The model further makes several empirically testable predictions about the failure dynamics, all of which are systematically verified across all three datasets. Second, I use COVID-19 as an example to study how policy and science respond to global emergencies. I find close co-evolution between COVID-19-related science and policy, where many policy documents in the COVID-19 pandemic substantially access recent, peer-reviewed, and high-impact science, and policy documents that cite science are especially highly cited within the policy domain. Yet at the same time, there is heterogeneity across policy-making institutions, where the tendency for policy documents to cite science appears mostly concentrated within intergovernmental organizations and much less so in national governments. Lastly, I study the public use and funding of science, by linking tens of millions of scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains—government documents, news media, and marketplace invention. I find public uses of science present a rich landscape of specialized consumption, yet, collectively, what the public uses and what scientists themselves use are closely consistent, and the funding of science closely tracks quantifiable public use, highlighting a remarkable alignment between scientific use, public use and funding.

Acknowledgements

I would like to express my deep gratitude to my advisor, Prof. Dashun Wang, for accepting me into his lab, offering me with his endless support, and helping me to grow both professionally and personally. Dashun is not only a perfect mentor who has offered me numerous guidance and opportunities for my research and career, but also a perfect role model who I can always look up to. I have learned so much from his enthusiasm, ambition, optimism, vision, and diligence, and feel so fortunate to have him as my advisor along this amazing journey. I am also grateful for my co-advisor, Prof. Noshir Contractor, for his insightful advice and suggestions since the first day of my graduate school. He is always willing to offer his help generously whenever needed, teaching me how to become an independent scholar. I would also like to thank Prof. Jorge Nocedal for serving on my prospectus and thesis committees. I have benefited from many exciting conversations with him.

The dissertation would not be possible without many wonderful collaborations with my coauthors and collaborators: Jichao Li, Santo Fortunato, Yang Wang, James Evans, Jian Gao, Ben Jones, Yuxiao Dong, Kuansan Wang, Kyle Myers, Wei Yang Tham, Karim Lakhani, Nina Cohodes, Jerry Thursby, Marie Thursby, Peter Schiffer, Joseph Walsh, Ryan Hill, and Carolyn Stein. I would especially thank James Evans and Ben Jones. They have taught me how to think like a social scientist and offered immense support in my career growth. Yang has provided a lot of help when I first joined the group.

All three studies presented in this work have improved tremendously thanks to external reviewers during the journal submission process, including Henry Sauermann, Shlomo Havlin, Carolin Haeussler, Paula Stephan, and many other anonymous ones.

I am proud to be a member of Northwestern Institute on Complex Systems and Kellogg Center for Science of Science and Innovation. I am grateful for the many past and current members of the family, especially Prof. Brian Uzzi, Prof. Hyejin Youn, Prof. Adam Pah, Ching Jin, Lu Liu, Zhongyang He, Wooseong Jo, Suman Maity, Nima Dehmamy, Yi Bu, Meijun Liu, Giorgio Tripodi, Minsu Park, Kariyushi Rao, Zander Furnas, Diego Gómez-Zará, Yifan Qian, Pramesh Singh, Yifang Ma, Yang Yang, Yuan Tian, Youyou Wu, Jacqueline Ng Lane, Joshua Becker, Sourav Medya, Kat Albrecht, Inho Hong, Hyunuk Kim, Frank van der Wouden, and Moh Hosseinioun. It is also my great pleasure to mentor several students in our lab, especially Zishan Gu, Zihang Lin and Zifeng Liu. Our time together has brought me numerous inspirations, excitement, and joy. Alanna Lazarowich, Krisztina Eleki, Meghan Stagl, and many others in the administrative team have worked tirelessly to keep the center running.

I am deeply indebted to many people in the IEMS department. I wish to thank IEMS faculty for admitting me into the PhD program six years ago, offering the unique opportunity for a computational social scientist to learn many valuable tools in operations research. Thanks to IEMS staff members, especially Jo Ann Yabolonka, Stephen Pederson, and Agnes Kaminski for their warmth and patience whenever I need help. Many of my friends in IEMS – Ruby Tu, Yuchen Xie, Yi Chen, Muchen Zhao, Yintai Ma, Liwei Zeng, Xin Qian, and Fengqiao Luo, just to name a few – have made the past few years a particular enjoyable time for me. Special thanks go to my cohort, roommate, and best friend Yuchen, who has always been knowledgeable, generous, and cheerful. I will surely miss the time we spent talking, eating, and laughing together.

I want to thank many of my friends whose existence has made my life incredibly bright and colorful. In particular, Xinzhou Ge, Minghe Liu, Yi Chen, Tingting Tang, Aibo Gong, and Ying Ni are always willing to offer their time and support, even across different time zones. Wenyuan Li has always been a caring, creative and optimistic friend and roommate.

I express my deepest gratitude to my mom Peifen Zhao and my dad Jianqiang Yin. Without their unconditional love and support over the many years, I can't imagine being even close to where I am now. I also want to thank my aunt Caifeng Yin, my grandparents, and others in my family for their continued love, care, and believe in me. Finally, I want to thank with love my girlfriend Binglu Wang, for her enduring love, accompany, encouragement, and support. She has always held me tight during all the ups and downs, brining countless happiness, inspiration, and strength to me. Thank you for being in my life.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The recent explosion of large-scale datasets in science and technology has offered an unprecedented opportunity to capture the entirety of scientific enterprise at a level of scale and detail that was previously unimaginable [1, 2]. This increasing availability of scholarly big data has further given rise to the emergence of the science of science and innovation, an interdisciplinary paradigm that leverages computational tools in complex systems and artificial intelligence to understand the precursors of impactful science and innovation. Building on canonical large-scale datasets such as Web of Science and Scopus, recent advances in this area has yielded a rich set of highly reproducible patterns underlying knowledge, careers, and collaboration, ranging from novel discoveries to scientific disciplines, from individual careers to team assemblies to environmental effects, from innovation policy to hiring and promotion to the assignment of prizes.

At the same time, the bulk of current research mostly builds on publication and citation databases, focusing on the ecosystem of *scientific publications* that got *published*. Yet science functions as a *multidimensional* complex adaptive system that extends beyond the published papers themselves. For example, while the existing insights from papers that got published are mostly limited to ideas, individuals, or teams that have succeeded in the first place, most innovations fail, sometimes in a speculative manner. This highlights one critical missing chapter in the field – failures. Indeed, given the widespread risk and uncertainty in scientific discoveries and technological inventions, our quantitative understanding of achievements and progress may be substantially biased and distorted unless we systematically look into failures. At the same time, as an integral part of modern society, science does not evolve on its own. Rather, it has long been described as a social institution that interacts with many other aspects of human society, which affects our ability to confront some of today's biggest challenges – from the pandemic to climate change, from fake news to privacy and security. Hence there is an urgent need to understand the role and impact of science outside science – in the halls of government, public perceptions, marketplace applications, and more.

Comprehensive investigations of these questions are only possible recently, thanks to the rapid development of novel approaches to data collection, linkage, and analysis. Building on these recent advances, in this dissertation I present how an interdisciplinary combination of (i) canonical theories from fields as diverse as history, sociology, psychology, and economics; (ii) large-scale datasets tracing various aspects and phases of science and innovation; and (iii) system modeling techniques rooted from complexity and network sciences can shed new light on these questions. Using millions of data on scientific publications, USPTO patents, funded research projects, policy documents, and mainstream news articles, these explorations not only uncover how various interconnected factors contribute to the progress and advancement in science, the fundamental engine of growth and prosperity, but also hold important implications for entrepreneurship and sustainable technological and business innovation.

The rest of this dissertation is organized as follows.

Chapter 2 presents an initial yet critical step towards our quantitative understanding of failures [3], the essential prerequisite to eventual success. To quantify the dynamics of failure, we have developed a simple one-parameter stochastic model to mimic how future attempts build on previous failures. The model makes four different empirically testable predictions, including one that is particularly surprising: Those who eventually succeed following failures and those who do not may be initially similar but are characterized by fundamentally distinct efficiency and quality trajectories, discernible long before the eventual outcome becomes apparent. We test the four predictions from datasets across science, startups, and security, finding broadly consistent empirical support across all three domains, which systematically verifies each prediction of our model. Together, these findings unveil identifiable yet previously unknown early signals that allow us to identify failure dynamics that will lead to ultimate success or failure.

Chapter 3 examines the coevolution between science and policy during the COVID-19 pandemic, asking a timely and important question: is our policy understanding closely linked to, or largely separated from, the evolving scientific understanding of the pandemic? Our analysis combines two largescale databases to capture policy, science, and their interactions and shows that policy documents in the pandemic substantially access recent, peer-reviewed, and high-impact science. At the same time, policy documents that cite science have especially high impact within the policy domain. Together, the close coevolution indicates a key link between policy and science is operating, which offers an important message for the scientific community, as scientists, journals, and funders work expeditiously to advance new research.

Building on these results, Chapter 4 develops a comprehensive framework to study how science is consumed in public domains. In particular, we e examine public use and public funding of science by linking tens of millions of scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains—government documents, news media, and marketplace invention. We find that different public domains draw from various scientific fields in specialized ways, showing diverse patterns of use. Yet, amidst these differences, we find universal alignment between what the public consumes and what is highly impactful within science. Further, a field's public funding is strikingly aligned with the field's collective public use. Overall, public uses of science present a rich landscape of specialized consumption, yet, collectively, science and society interface with remarkable alignment between scientific use, public use, and funding.

CHAPTER 2

Quantifying the dynamics of failure

Henry Ford went bankrupt five times before founding the Ford Motor; J.K. Rowling was rejected by twelve publishers before introducing Harry Potter to the world; Yet neither came close to Thomas Edison, who famously failed more than one thousand times before identifying the carbon filament for light bulb. Human achievements are often preceded by repeated attempts that fail, yet little is known about the mechanisms governing the dynamics of failure, due to lack of empirical data sources as well as theoretical foundations for systematic quantification of failures. Indeed, Thomas Edison once said, '*Many of life's failures are people who did not realize how close they were to success when they gave up.*' Yet as we show in this chapter, by modeling the process of repeated failures as a stochastic complex system, there are rich identifiable early signals that can help us predict the eventual outcome to which failures lead.

2.1. Data description

One important empirical challenge in understanding and modeling failures is the lack of ground-truth information that contains unbiased records for both successful and failed attempts. To understand the dynamics of failure, we collected three large-scale datasets.

Here we compiled a comprehensive database consisting of three large-scale datasets across three different domains: Dataset D_1 contains submission histories of individual scientists in the US National Institutes of Health (NIH) grant system. D_2 contains profiles of innovators together with their startup ventures recorded in the VentureXpert investment database. D_3 records terrorist organizations and attacks retrieved from the Global Terrorism Database.

2.1.1. NIH grant application dataset

The first dataset (D_1) contains all R01 grant applications (776,721 in total) that have been ever submitted by 139,091 scientists to NIH from 1985 to 2015. For each grant application, we obtained its evaluation score (if reviewed on a panel), a unique identifier for the PI, PI name and the application outcome (funded/not funded), allowing us to reconstruct individual application histories and their repeated attempts to obtain funding.

The NIH grant application dataset represents an excellent setting to study dynamics of failure for several reasons. First, it contains ground-truth information for both successes and failures. Second, as the world's largest public funder for biomedical research, NIH is the dominant funding source for biomedical scientists in the US [4, 5]. Indeed we tracked funding acknowledgment information cited within biomedical research papers, finding among all PubMed papers published in the US (2008 to 2015), NIH represents the majority of funding sources (81% out of top 10 agencies). R01 is the most common research funding mechanism within the NIH [4, 6, 7], accounting for the majority of the total funding. To compare the dynamical pattern between R01 and other granting mechanisms, we downloaded successful NIH grants from other mechanisms from NIH Research Portfolio Online Reporting Tools (RePORT), finding R01 grants are uniformly distributed within all NIH grants one obtains throughout a career. Here we extract all new grant applications (excluding renewals, revisions and resubmissions) to reconstruct sequences of attempts. We truncate each sequence if (i) the individual gets one grant (successful group); or (ii) the individual has been inactive for a long period (unsuccessful group).

2.1.2. VentureXpert investment dataset

Our second dataset (D_2) traces start-up investment records from the VentureXpert (SDC Platinum) database, including 58,111 startup companies and 163,106 investment rounds from 1970 to 2016. Tracing every startup in which VCs invested, D_2 allows us to reconstruct individual career histories counting successive ventures in which they were involved. For each investment we obtained information on investment amount, funding date, company name and a full list of innovators involved. We then link these records with company information on Initial Public Offering and Merger & Acquisitions as outcome variables. Following the entrepreneurship literature [8–10], we match individual entrepreneurs and startup ventures by linking each company with people listed as executives or board members at the first funding round. One advantage of this dataset is that 98.7% records have complete information of first and last names rather than initials, allowing us to construct career trajectories of 253,579 innovators.

Among the existing datasets capturing startups, the VentureXpert database, the official database of the National Venture Capital Association is among the most comprehensive and authoritative databases [11]. To further explore the coverage of the database, we compare the number of IPOs within our data versus US total counts, finding our dataset captures a significant fractions of IPOs, with the ratio between the two statistics remaining stable over time, documenting the reliability of this dataset. We also cross-validated individual entrepreneurs coverage with Crunchbase. We select top 1000 serial executives and board members ranked by the number of different jobs in Crunchbase, finding more than 70% of the profiles are included in VentureXpert.

Another challenge in modeling dynamics of failure in startup datasets is the ambiguity of 'failures' [12], which could include bankruptcy, termination to prevent future losses, and deviation from desired results. Recognizing the complexity of this issue, here we closely follow existing literature on venture capital and serial entrepreneurship [9, 10]. We focus on all portfolio companies that have received at least one round of funding, and define those who went public or got acquired or merged at high values (percentile as compared with all M&As in the same year) as successes. We performed different measurement variations by changing the percentile threshold (1% and 5%) and also by only including IPOs. We find our results remain the same. If a company obtained its first investment but did not succeed within a certain period, this venture is marked as a failure. In this dataset we treat each new venture as an attempt, starting at the date of first round investment. Similar to D_1, sequences of attempts by each individual are collected into a sequence, where the stopping criterion is defined by either (i) the individual is involved in one company that eventually achieved IPO or high-value M&As (successful group); or (ii) the individual has been inactive for a long period without success (unsuccessful group).

2.1.3. GTD terrorism attack dataset

Going beyond traditional innovation domains, we collected our third dataset (D_3). D_3 contains 170,350 terrorist attacks by 3,178 organizations from 1970 to 2017, collected by the Global Terrorism Database, one of the most systematic databases on domestic and transnational terrorist events [13]. For each attack we obtain information on its date, type, location, and consequences in terms of the number of people killed and wounded. Some records in this corpus are based on speculation or dubious claims of responsibility, which are discarded in our analysis to ensure the data quality.

There lacks a clear definition of 'success' for terrorist attacks, partly due to their diverse intents and consequences. To be consistent with our empirical steps in D_1 and D_2, here we treat an attack as successful if it killed at least one victim. To this end, we collected sequences of attacks of each terrorist organization, and classify the samples as (i) the organization killed at least one people (successful group); or (ii) the organization has been inactive for a long period without success (unsuccessful group).

One potential concern with this definition is that goals of terrorist attacks differ, and not all attacks are aimed at killing victims. This concern is somewhat alleviated since (1) 84.7% the attacks were targeted at human beings (i.e. assassination, bombing/explosion and assault) and (2) human-targeted attacks were uniformly distributed within full attack history of terrorist organizations. To rule out the possibility that samples in unsuccessful group are simply those who do not aim for killing victims, we further remove samples from the unsuccessful group if more than half of the events in this sample are not human-targeted. We also performed robustness checks by performing the same operation on successful group or using the full sample in unsuccessful group, finding our results remain robust. Although these checks do not necessarily account for the diverse goals of terrorist attacks, they do consistently show no evidence of systematic bias.

2.2. Modeling the dynamics of failure

Building on the rich literature about innovation [1, 14–19], human dynamics [20–23] and learning [24–29], we develop a simple one-parameter model that mimics how successful future attempts build on those past.

2.2.1. Mechanisms of chance and learning

Chance and learning [25, 28] are two primary mechanisms explaining how failures may lead to success. If each attempt has a certain likelihood of success, the probability that multiple attempts all lead to failure decreases exponentially with each trial. The chance model therefore emphasizes the role of luck, suggesting that success eventually arises from an accumulation of independent trials. To test this, we systematically compare the performance of the first and penultimate attempt within failure streaks, measured by NIH percentile score for a grant application (D_1), investment size by VCs to a company (D_2), and number of wounded individuals by an attack (D_3), as detailed below.

For the NIH grant application dataset, we make use of the percentile scores assigned by NIH review panels. NIH uses a two-step peer review mechanism: Roughly half of the proposals are selected for the second round discussion, where each proposal is given a percentile score based on their percentile ranking among its peers. Percentile score has been widely used to measure the quality of R01 grant applications [7, 30], reflecting judgment of expert reviewers. Although reviewers score are necessarily imperfect, there is growing evidence for strong correlations between percentile score and subsequent successes of the project [5, 31]. One disadvantage of using the percentile score is that undiscussed proposals (those get rejected in the first round) do not have such scores. Moreover, since there exist differences concerning the discussion rate, applications lying on the boundary of discussion can have either marginal scores or no scores. Indeed, here we calculate the proportion of having a percentile score around 57% and plot the score distribution. We find as score exceeds 50, there are much fewer samples, since many proposals at this rank did not even get discussed and assigned a score. To avoid discrimination across study sections, here we take score below 50 and regard the remaining proposals as undiscussed. We also vary the threshold to 55, finding results remain the same. Lower percentile scores indicate better performance. To be consistent with other measures (higher the better) we rescale the percentile scores using 1-0.01×original score, so the values reported in main text are bigger the better.

To measure the performance in startup ventures, we leverage the investment amount in the first funding round as a proxy. Although there are a series of firm-level statistics that could potentially measure the quality of a venture, investment amount stands out as a preferred choice of representing investor evaluations. This definition does not account for geographical and industrial factors, as such information is not available to us, but it serves as a reasonable index of startup companies potentials in achieving their eventual goals (IPO or high-value M&As).

Similar to other frequently used measures in economics, investment amount follows a fat-tailed distribution and exhibits time-dependent properties. To address the two challenges, we take logarithmic of the investment amount and calculate z-score within each year. Denoting the amount of all investments made in year t as $\{s_1^t, \cdots, s_n^t\}$, here we rescale the values into the performance score z through

$$z_i^t = \frac{\log(s_i^t) - \mathrm{E}[\log(s^t)]}{\sqrt{\mathrm{Var}[\log(s^t)]}}$$

Once rescaled, we find z_i^t approximately follow the standard normal distribution $N(0, 1)$ independent from t, allowing us to directly compare attempts made in different years. We then compare first-round investment amounts for successful and failed attempts, finding the two samples are clearly separated.

Similarly, for terrorist attacks, one measure for performance is the number of individuals wounded, which is reported for more than 91% of the attacks recorded in the database. To this end, we collect wound statistics as our performance measure. Indeed, fatal (successful) attacks also lead to a higher number of wounded individuals than others, validating the effectiveness of using wounded statistics as performance measurements. Related studies of terrorist attacks suggest the outcome of attacks follow a power law distribution, which is also confirmed in our dataset. To this end, we rescale the original values by $\log(wounded + 1)$ in our analysis.

Note that although the overall coverage of performance measures is high (94% for D_2 and 91% for D_3), in both datasets there are missing values. To ensure that they do not affect our results, we also label these missing values as NA and exclude them as we analyze performance dynamics. Analyses that do not require performance information are measured on the full data sample.

We find that across all three datasets, the penultimate attempt shows systematically better performance than the initial attempt (Figs. 2.1c-e). These results reject that success is simply driven by chance (Fig. 2.1a) but lend support to the learning mechanism

(Fig. 2.1b), which suggests that failure may teach valuable lessons difficult to learn otherwise [24, 25, 28]. As such, learning reduces the number of failures required to achieve success, and predicts that failure streaks should follow a narrower length distribution (Fig. 2.1g) than the exponential one predicted by chance (Fig. 2.1f).

To this end, we empirically measure the distribution of failure streak length, defined as the number of failures before success. Across all three domains, failure streak length follows a fat-tailed distribution (Figs. 2.1h-j), indicating that despite performance improvement, failures are characterized by longer-than-expected streaks prior to the onset of success. To further test these results, we perform two randomization processes. We performed our first randomization operation, by keeping the timing and outcome of each attempt but changing the individual/organization associated with the attempt via random selection. The null model leads to exponentially distributed failure streaks. We then performed a second randomization procedure by taking the samples used in Fig. 2.1 and shuffling the success/failure label from each attempt. This operation keeps constant both the overall success rate and the total number of attempts for each individual (Fig. 2.7c-e). The two versions of randomization both lead to exponential like distributions, showing clear deviation from data.

Together, these observations demonstrate that neither chance nor learning alone can explain the empirical patterns underlying failures, suggesting that more complex dynamics may be at work.

Note that Fig. 2.1h-j and 2.4a-c only show results for less than 21 consecutive failures prior to the eventual outcome, accounting for 99.99%, 100%, 99.35% for the successful

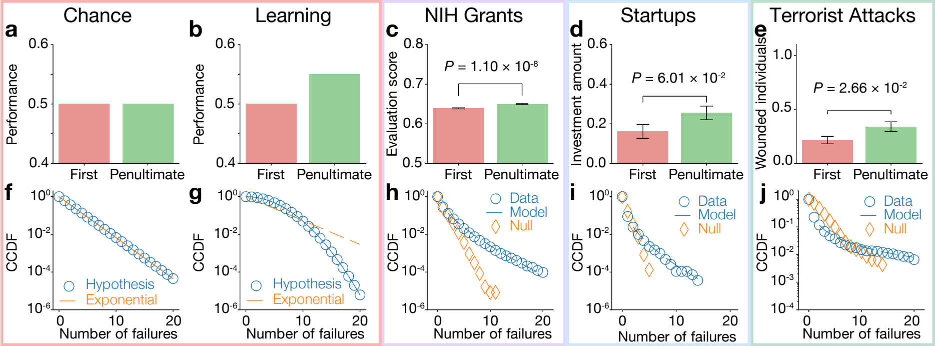

Figure 2.1. **Mechanisms of chance and learning.**

We compare theoretical predictions and empirical measurements for performance changes (**a-e**) as well as the length distribution of failure streaks (**f-j**). The chance model predicts no performance change (**a**), with failure streak length following an exponential distribution (**f**). The learning hypothesis predicts improved performance (**b**), with shorter failure streaks than expected by the chance model, corresponding to a faster-than-exponential distribution (**g**). Both hypotheses are contested by empirical patterns observed across the three datasets. To ensure that performance metrics are comparable across data and models, we standardized performance measures according to their underlying distribution. We find that failures in real data are associated with improved performance between the first and penultimate attempt. Yet at the same time, failure streaks are characterized by a fat-tailed length distribution, indicating that failure streaks in real data are longer than expected by chance (**h-j**). For clarity, here we show results for failure streaks whose length is less than 21. We further construct a randomized sequence of successes and failures by assigning each attempt to agents at random. We find that failure streak length in randomized sequence follows an exponential like distribution, showing clear deviations from data.

group and 99.99%, 100%, 99.60% for the unsuccessful group. All statistical tests are performed on the full data (100%).

2.2.2. The k model

In order to formulate a new attempt, the individual needs to go through every component, and decide what to do next. For a past attempt j, each component i is characterized by an evaluation score $x_j^{(i)}$, which falls between 0 and 1. The agent can either create a new version (with probability p), or with probability $1-p$ reuse an old one by choosing among past versions. The main cost of creating a new version is time. Here we assume each new version takes one unit of time, and upon creation takes up an evaluation score, drawn randomly from a fixed distribution $\rho(x)$. Real systems are likely to differ in their specific score distributions. Here for simplicity, we assume $\rho(x)$ follows a uniform distribution on $[0, 1]$, approximating the percentile of any underlying score distributions real systems may follow. One difference between our model and canonical learning curve models [32] is that one has little information on the new versions until it gets implemented and evaluated, hence new versions are not guaranteed to increase or decrease their score.

Of the many factors that may influence p, one key factor is the quality of existing versions. Denoting with x^* the best score among past versions, we expect p to be a function of x^*. Indeed, consider the two extreme cases. If $x^* \to 0$, existing versions of this component have among the worst scores hence a high potential to be improved upon with a new version. Therefore the likelihood of creating a new version is high, i.e., $p \to 1$. On the other hand, $x^* \to 1$ indicates an already excellent version, corresponding to a decreased incentive to create a new one ($p \to 0$). Reusing the existing best version allows the particular component to retain its score x^* and also avoids incurring additional time cost the individual can avoid spending time working on. To this end, considering $P(x \geq x^*) = 1 - x^*$ as the potential to improve on existing versions, we assume $p = (1 - x^*)^\alpha$, where $\alpha > 0$ characterizes an individual's propensity to create new versions given the quality of existing versions. The higher this potential, the more likely one may create a new version [33].

The dynamics of quality score, x_n, can be captured by a higher-order Markov process of memory length k, following

$$x_n^* = \max\{x_{n-k}, \cdots, x_{n-1}\} \tag{2.1}$$

$$x_n \sim \begin{cases} U[0,1], \ w.p. \ (1-x_n^*)^\alpha \\ \delta(x-x_n^*), \ w.p. \ 1-(1-x_n^*)^\alpha \end{cases} \tag{2.2}$$

where we assume $x_n = 0$ for all $n < 0$.

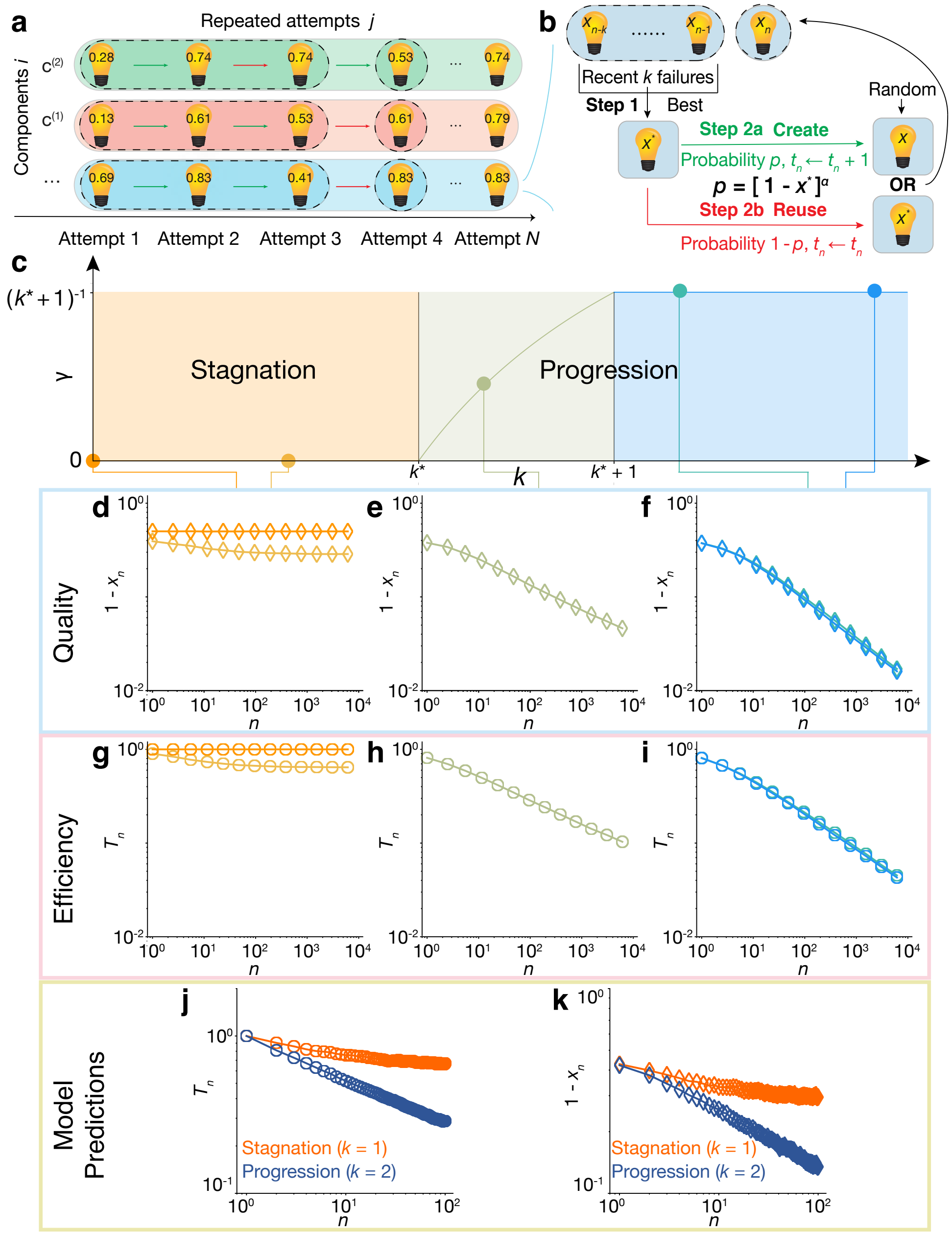

Figure 2.2. **The k model.**

(Caption next page.)

(Previous page.) (**a**) We treat each attempt as a combination of many independent components. For attempt j, each component i is characterized by an evaluation score $x_j^{(i)}$. The score for a new version is often unknown until attempted, hence a new version is assigned a score, drawn randomly from $[0, 1]$. (**b**) To formulate a new attempt, one can either create a new version (with probability p, green arrow), or reuse an existing version by choosing the best one among past versions x^* (with probability $1 - p$, red arrow). Indeed, $P(x \geq x^*) = 1 - x^*$ captures the potential to improve on prior versions, prompting us to assume $p = (1 - x^*)^\alpha$. **(c)** Analytical solution of the model reveals that the system is separated into three regimes by two critical points k^* and $k^* + 1$. The solid line shows an extended solution space of our analytical results. **(d-i)** Simulation results from the model ($\alpha = 0.6$) for quality (d-f) and efficiency (g-i) trajectories for different k parameters, showing distinct dynamical behavior in different regimes. All results are based on simulations averaged over 10^4 times. **(j,k)** Phase transition around k^* predicts the coexistence of two groups that fall in the stagnation and progression regimes, respectively.

The parameter k in our model can be viewed as approximating the 'memory' of past versions. The rationale of using k for the model is rooted in the learning literature, showing that the general notion of 'forgetting' takes multiple forms, often representing a combination of individual, organizational and environmental factors. Indeed, several relevant factors may be at play, which can generate patterns similar to 'forgetting'. For example, in rapidly shifting innovation domains, not all past failures remain useful over time, and some become obsolete. Consider the concept of 'knowledge depreciation' [34], which could also apply in our settings as environments (scientific knowledge/capital markets/security situations) evolve over time, such that past experience could become useless even if memorized. For example, an NIH proposal four failures ago may become irrelevant as the ideas proposed have been dispositively proven wrong, or published by the PI or another research group [35, 36]. Similarly, startup ideas from the dot com era may be irrelevant in the era of AI and Blockchain [9]. Terrorist tactics can also depreciate over time, as past strategies attracted media coverage and gave rise to tighter security measures defending against them [37]. This line of reasoning supports the idea that recent attempts are most relevant. It is also consistent with the learning literature, which suggests knowledge ‘forgetting’ can happen in distinct ways, either voluntarily or involuntarily [38]. Motivated by these reasons, here we select a single parameter k to encapsulate a variety of potential contributing factors.

2.2.3. Two extremes of the model

Next, we start with two extreme cases of the model. $k = 0$ means each attempt is independent from those past. In this case, one creates a new version every time, hence for all n we have

$$x_n \sim U[0,1] \tag{2.3}$$

and

$$t_n \equiv 1 \tag{2.4}$$

Here our model recovers the chance model, predicting that as n increases, both $\langle x_n \rangle$ and $\langle t_n \rangle$ remain constant (Fig. 2.3ad). That is, without considering past experience, failure does not lead to quality improvement. Nor is it more efficient to try again.

The other extreme ($k \to \infty$) considers all past attempts. e can rewrite the process as

$$x_n^* = \max\{x_0, \cdots, x_{n-1}\} \tag{2.5}$$

$$x_n \sim \begin{cases} U[0,1], \ w.p. \ (1-x_n^*)^\alpha \\ \delta(x-x_n^*), \ w.p. \ 1-(1-x_n^*)^\alpha \end{cases} \tag{2.6}$$

Here we focus on the dynamics of x^*, obtaining

$$x_{n+1}^* \sim \begin{cases} U[x_n^*,1], \ w.p. \ (1-x_n^*)^{\alpha+1} \\ \delta(x-x_n^*), \ w.p. \ 1-(1-x_n^*)^{\alpha+1} \end{cases} \tag{2.7}$$

where $x_1^* \sim U[0,1]$. To this end, let us denote f_n as the probability density function of x_n^*, obtaining

$$f_{n+1}(x) = f_n(x)(1-(1-x)^{\alpha+1}) + \int_0^x f_n(y)(1-y)^\alpha dy \tag{2.8}$$

with $f_1(x) \equiv 1$ for $x \in [0,1]$. By induction we obtain

$$f_n(x) \sim [1-(1-x)^{\alpha+1}]^{n-1} \tag{2.9}$$

Therefore we have

$$\begin{aligned} t_n &= \frac{\int_0^1 (1-x)^\alpha f_n(x)dx}{\int_0^1 f_n(x)dx} \\ &= \frac{B(n,1)}{B(n,1/(\alpha+1))} \\ &\sim \Gamma\left(\frac{1}{\alpha+1}\right)^{-1} n^{-\frac{\alpha}{\alpha+1}} \end{aligned} \tag{2.10}$$

and

$$
\begin{aligned}
1 - x_n &= \frac{\int_0^1 \{(1-x)[1-(1-x)^\alpha)] + (1-x)^\alpha/2\} f_n(x)dx}{\int_0^1 f_n(x)dx} \\
&= \frac{B(n, 2/(\alpha+1)) - B(n, 1+1/(\alpha+1)) + B(n,1)/2}{B(n, 1/(\alpha+1))} \\
&\sim \Gamma\left(\frac{1+\min\{\alpha,1\}}{\alpha+1}\right)\Gamma\left(\frac{1}{\alpha+1}\right)^{-1} n^{-\frac{\min\{\alpha,1\}}{\alpha+1}}
\end{aligned}
\tag{2.11}
$$

The model therefore predicts a temporal scaling in failure dynamics. That is, the time it takes to formulate a new attempt decays with n, asymptotically following a power law (Fig. 2.3e):

$$
T_n \equiv \langle t_n \rangle / \langle t_1 \rangle \sim n^{-\gamma}, \tag{2.12}
$$

where $\gamma = \gamma_\infty = \alpha/(\alpha+1)$ falls between 0 and 1. Besides increased efficiency, new attempts also improve in quality, as the average potential for improvement decays following $\langle 1 - x_n \rangle \sim n^{-\eta_\infty}$, where $\eta_\infty = \min\{\gamma_\infty, 1 - \gamma_\infty\}$ (Fig. 2.3b). Here the model recovers the canonical result from the learning literature [24, 27, 39–41], commonly known as Wright's Law [42]. This is because, as experience accumulates, high-quality versions are preferentially retained, while their lower quality counterparts are more likely to receive updates. As fresh attempts improve in quality (Fig. 2.3b), they reduce the need to start anew, thus increasing the efficiency of future attempts (Fig. 2.3e).

2.2.4. Solving the general model

These two limiting cases might lead one to suspect a gradual emergence of scaling behavior as we learn from more failures. Yet, here we show that, as one increases parameter k, the

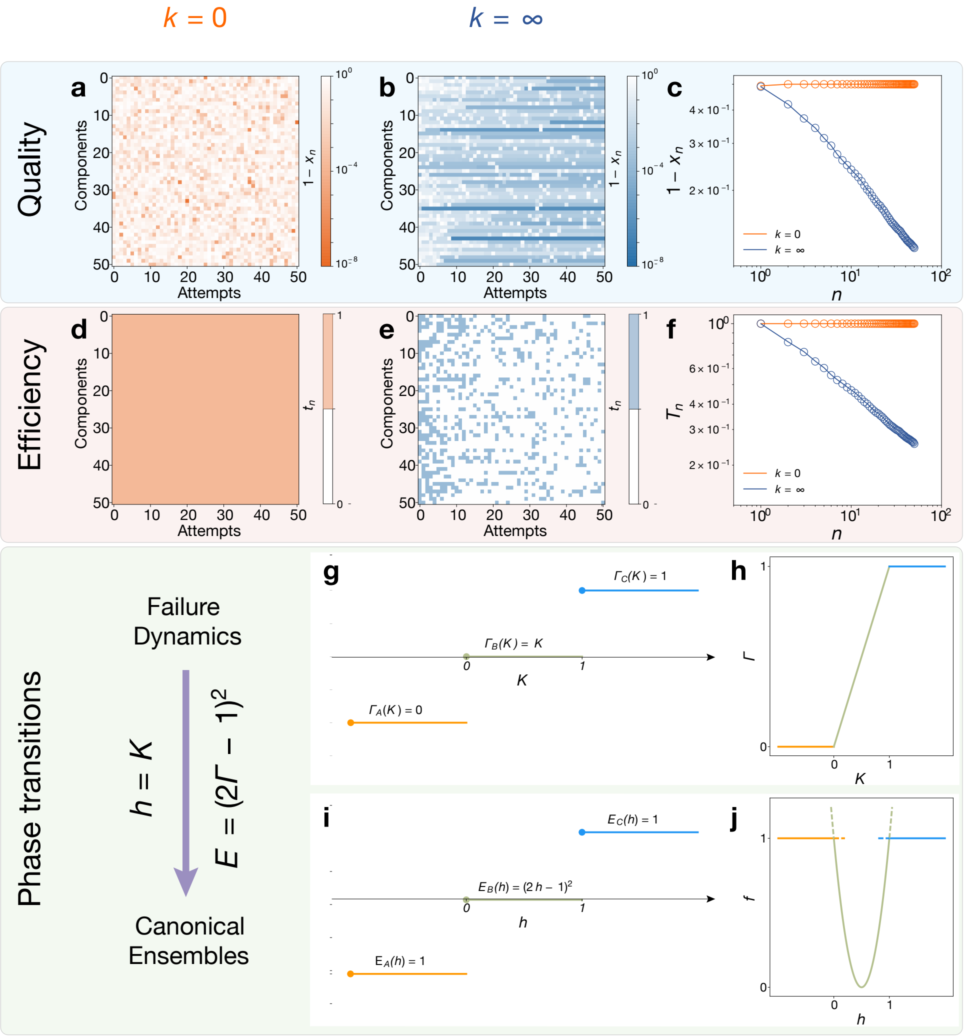

Figure 2.3. **Understanding the k model.**

(Caption next page.)

scaling exponent γ follows a discontinuous pattern (Fig. 2.2c) and only varies within a narrow interval of $\lfloor k^* \rfloor < k < \lceil k^* \rceil + 1$ $(k^* \equiv 1/\alpha)$.

To solve the k-model with higher-order dependency, here we use a series of careful approximation techniques as well as results from renewal process theories [43]. More

(Previous page.) **a-f**, Simulation results from the model ($\alpha = 0.6$) for the cases of $k = 0$ (**a,d**) and $k \to \infty$ (**b,e**) in terms of the average quality (**a-c**) and efficiency (**d-f**) of each attempt. $k = 0$ recovers the chance model, predicting a constant quality (**c**) and efficiency (**f**). $k \to \infty$ predicts temporal scaling that characterizes the dynamics of failure (**g**) with improved quality (**d**), recovering predictions from learning curves and Wright's Law. **g-j**, Illustration of mapping between failure dynamics (**g,h**) and canonical ensembles (**i,j**). The canonical system is characterized by three different states a, b, c with corresponding energy density $E_a(h), E_b(h), E_c(h)$. Here we assume $E_a(h) = (2\epsilon h - 1)^2$, $E_b(h) = (2h - 1)^2$, and $E_c(h) = [2\epsilon(1-h)-1]^2$, where $\epsilon \to 0^+$. The introduction of ϵ is to distinguish state a from state c, both of which can be approximated in the limiting condition $E_a(h) = E_c(h) = 0$. We map $f \to (2\Gamma - 1)^2$, $N \to \ln n$, $h \to K$, and $E_i(h) = [2\Gamma_i(K) - 1]^2$. In this case, the two transition points k^* and $k^* + 1$ correspond to $h = 0$ and 1 in the canonical ensemble systems.

specifically, we first note that

$$|\{n_1 \leq n \leq n_2 : x_n = x_m^*\}| \leq n_2 - n_1 + 1 \tag{2.13}$$

$$|\{n_1 \leq n \leq n_2 : x_n = x_m^*\}| \geq \sum_{i=0}^{[(n_2-n_1)/k]-1} \sum_{j=0}^{k-1} I(x_{n_1+ki+j} = x^*_{n_1+ki+i}) \geq [(n_2 - n_1)/k] \tag{2.14}$$

Hence to calculate the length of a sequence, we only need to estimate the number of versions that are once baseline versions (i.e. n such that $x_n = x_m^*$ for some $n + 1 \leq m \leq n + k$).

Denote $z_m = 1 - x_n^*$ as all such baseline scores. We now calculate for a specific z_m to be taken by a new one, the number of attempts it takes. Indeed, given a score z_m and assuming that it has been reused as $z_m = z_{m-1}$, we have

$$z_{m+1} = \begin{cases} z_m & w.p.\ \frac{[1-z_m^{k\alpha}(1-z_m)^k](1-z_m^\alpha)}{1-z_m^\alpha(1-z_m)} \sim O(1) \\ U[0, z_m] & w.p.\ \frac{[1-z_m^{k\alpha}(1-z_m)^k]z_m^{\alpha+1}}{1-z_m^\alpha(1-z_m)} \sim O(z_m^{\alpha+1}) \\ \min\{U_1[0,1], \cdots, U_k[0,1]\} & w.p.\ z_m^{k\alpha}(1-z_m)^k \sim O(z_m^{k\alpha}) \end{cases} \tag{2.15}$$

Here we use the big-O notation to find the asymptotic case for $z_m \to 0$. This equation shows two important insights:

(1) If we calculate the number of iterations that z_m gets reused, it should be in the order of $O(z_m^{-\min\{k\alpha,\alpha+1\}})$, leading to two cases that will be discussed in detail.

(2) There exist two different forces for the substitution of baseline versions to happen: quality (with probability $O(z^{k\alpha})$) and recency (with probability $O(z^{\alpha+1})$). For $k\alpha < \alpha+1$, the recency mechanism dominates for small z, i.e. produces a worse succeeding score. Hence, it keeps a stable score distribution of new baseline scores as n increases. However, once $k\alpha > \alpha + 1$, quality mechanism takes over for small z, characterizing a continuous path of improvement.

Here, we first derive our results for the regime $k\alpha < \alpha + 1$, and then extend the obtained results to the other regime.

2.2.4.1. Case 1: $k\alpha < \alpha + 1$**.** When $z_{m+1} \neq z_m$, our previous results show that with high probability, z_m is the extreme value among k i.i.d. random variables on $U[0,1]$, hence the pdf of z_m, $f(z_m) \sim$ *const* as $z_m \to 0$. Below we offer a more rigorous proof: Take all the different z_m as $\tilde{z}$ and consider a limiting distribution of $f(\tilde{z})$. We have

$$f(\tilde{z}) \sim \int_0^1 f(\tilde{z}')O(1)d\tilde{z}' + \int_{\tilde{z}}^1 f(\tilde{z}')O(\tilde{z}'^{\alpha+1-k\alpha})/\tilde{z}'d\tilde{z}' \tag{2.16}$$

Assuming $f(\tilde{z}) \sim \tilde{z}^{\beta_1}$ and consider $\tilde{z} \to 0$ one gets

$$\beta_1 = \min\{0, 1, \beta_1 + \alpha + 1 - k\alpha\} = \min\{0, \beta_1 + \alpha + 1 - k\alpha\} \tag{2.17}$$

Since $k\alpha < \alpha + 1$, we get $\beta_1 = 0$. Hence, as we generate a new baseline score satisfying $z_m \neq z_{m-1}$, we approximate the number of iterations it will be retained as $u \sim z^{-k\alpha}$. Let $z_m = z_{m+1} = \cdots = z_{m+u}$. For z_{m+u+1} we take a new random variable from a fixed distribution on $[0, 1]$ whose probability density does not diverge near 0. If we consider the change of baseline scores as a 'jump' and number of iterations of repeated reuse as the length of this jump (u), we eventually arrive at a Levy flight [44].

We can define $u_i \equiv z_i^{-k\alpha}$, following asymptotically power law pdf $P(u) \sim u^{-1/k\alpha-1} \equiv u^{-\mu-1}$, and $m(N) \equiv \min_m\{u_1 + \cdots u_m \geq N\}$. Next we solve $\langle u^\lambda_{m(N)} \rangle$ for some λ. We first calculate $P(u_{m(N)})$, which equals to

$$\begin{aligned} P(u_{m(N)} = u) &= P(u) \int_{\max\{N-u,0\}}^{N} \sum_{k=0}^{\infty} P_k(v)dv \\ &= P(u) \int_{\max\{N-u,0\}}^{N} G(v)dv \end{aligned} \tag{2.18}$$

where $P_k(v) \equiv P(v_1 + \cdots + v_k = v)$ and $G(v) \equiv \sum_{k=0}^{\infty} P_k(v)$. P_k can be obtained analytically by induction, following

$$P_k = \begin{cases} P_{k-1} \circ P, & k \leq 1 \\ \delta(0), & k = 0 \end{cases} \tag{2.19}$$

Hence we have

$$G = \sum_{k=0}^{\infty} P_k = G \circ P + \delta(0) \tag{2.20}$$

Taking the Laplace transformation we obtain

$$\tilde{G} = \frac{1}{1 - \tilde{P}} \tag{2.21}$$

The quantity of interest, $M(N) \equiv \langle u_{m(N)}^{\lambda} \rangle$, can be formulated as

$$\begin{aligned} M(N) &= \int_0^{\infty} P(u_{m(N)} = u) u^{\lambda} \\ &= \int_0^N P(u) u^{\lambda} \int_{N-u}^N G(v) dv du + \int_N^{\infty} P(u) u^{\lambda} \int_0^N G(v) dv du \\ &= \int_0^N Q(u)[H(N) - H(N-u)] du + \int_N^{\infty} Q(u) H(N) du \\ &= H(N) \int_0^{\infty} Q(u) du - \int_0^N Q(u) H(N-u) du \\ &= H(N) \int_0^{\infty} Q(u) du - (Q \circ H)(N) \end{aligned} \tag{2.22}$$

where $H(N) = \int_0^N G(v)dv$ and $Q(u) = u^{\lambda} P(u)$. Performing again the Laplace transformation, we obtain

$$\begin{aligned} \tilde{M} &= \tilde{H}(\int_0^{\infty} Q(u) du - \tilde{Q}) \\ &= \tilde{G}(\int_0^{\infty} Q(u) du - \tilde{Q})/s \\ &= \frac{\int_0^{\infty} Q(u) du - \tilde{Q}}{s(1 - \tilde{P})} \end{aligned} \tag{2.23}$$

Assuming

$$P(x) = \mu x^{-\mu-1} I(x \geq 1) \tag{2.24}$$

we obtain

$$\tilde{P}(s) = \mu s^{\mu} \Gamma(-\mu, s) \tag{2.25}$$

$$\tilde{Q}(s) = \mu s^{\mu-\lambda} \Gamma(\lambda - \mu, s) \tag{2.26}$$

$$\int_0^\infty Q(u) du = \frac{\mu}{\mu - \lambda} \tag{2.27}$$

where $\Gamma(a, s) = \int_s^\infty t^{a-1} e^{-t} dt$ is the upper incomplete Gamma function. Inserting these results into the previous function we arrive at

$$\tilde{M} = \frac{\mu/(\mu - \lambda) - \mu s^{\mu-\lambda} \Gamma(\lambda - \mu, s)}{s[1 - \mu s^{\mu} \Gamma(-\mu, s)]} \tag{2.28}$$

To obtain asymptotic results for $M(N)$ as $N \to \infty$, we approximate $\tilde{M}(s)$ as $s \to 0^+$. Here we use the following expansion

$$\Gamma(a, s) = \Gamma(a) - \frac{s^a}{a} + \frac{s^{a+1}}{a+1} + O(s^{a+2}) \tag{2.29}$$

The previous equation hence writes

(2.30)

$$\begin{aligned}\tilde{M} &\approx \frac{\mu/(\mu-\lambda) - \mu s^{\mu-\lambda}\Gamma(\lambda-\mu) + \mu s^{\mu-\lambda} s^{\lambda-\mu}/(\lambda-\mu) - \mu s^{\mu-\lambda} s^{\lambda-\mu+1}/(\lambda-\mu+1)}{s[1-\mu s^{\mu}\Gamma(-\mu) + \mu s^{\mu} s^{-\mu}/(-\mu) - \mu s^{\mu} s^{-\mu+1}/(1-\mu)]} \\ &= \frac{-\mu s^{\mu-\lambda}\Gamma(\lambda-\mu) - \mu s/(\lambda-\mu+1)}{s[-\mu s^{\mu}\Gamma(-\mu) - \mu s^{\mu} s^{-\mu+1}/(1-\mu)]} \\ &= \frac{s^{\mu-\lambda}\Gamma(\lambda-\mu) + s/(\lambda-\mu+1)}{s[s^{\mu}\Gamma(-\mu) + s/(1-\mu)]} \sim s^{\min\{\mu-\lambda,1\}-\min\{\mu,1\}-1}\end{aligned}$$

Hence we obtain

$$M = L^{-1}(\tilde{M}) \sim n^{-\min\{\mu-\lambda,1\}+\min\{\mu,1\}} \tag{2.31}$$

Let us consider the two specifications:

Case 1: $\lambda = -1/k$, we have $M \sim n^{\min\{1/(k\alpha),1\}-1}$, hence

$$\langle (1-x^*)^{\alpha} \rangle \approx M = \begin{cases} \text{const.}, & k\alpha \le 1 \\ n^{-1+1/(k\alpha)}, & k\alpha > 1 \end{cases} \tag{2.32}$$

Case 2: $\lambda = -1/(k\alpha)$, we have $M \sim n^{\min\{1/(k\alpha),1\}-\min\{2/(k\alpha),1\}}$, hence

$$\langle 1-x^* \rangle \approx M = \begin{cases} \text{const.}, & k\alpha \le 1 \\ n^{-1+1/(k\alpha)}, & 1 < k\alpha \le 2 \\ n^{-1/(k\alpha)}, & k\alpha > 2 \end{cases} \tag{2.33}$$

This eventually leads to

$$\langle 1-x \rangle = \langle z \rangle = \langle z^* + z^{*\alpha}/2 - z^{*(\alpha+1)} \rangle \approx \langle z^* + z^{*\alpha}/2 \rangle \sim n^{-\min\{\gamma,1-\gamma\}} \tag{2.34}$$

2.2.4.2. Case 2: $k\alpha > \alpha + 1$. As we discussed, in this regime the quality dynamics is dominated by the second mechanism, which does not depend on k, and asymptotically follows the same mechanism as learning from all failures model ($k = \infty$). Indeed, if we expand our solution and take $k \to (1 + 1/\alpha)^-$, we obtain $\gamma = 1 - 1/(k\alpha) \to \alpha/(\alpha + 1)$ and $\eta = \min\{\gamma, 1 - \gamma\} \to \min\{1, \alpha\}/(\alpha + 1)$, which are the same as $k = \infty$. Hence, the regime lying between $k = 1 + 1/\alpha$ and $k = \infty$ should have the same scaling behaviors.

Taken together, we obtain

$$\gamma = \begin{cases} 0, & k < k^* \\ 1 - k^*/k, & k^* \leq k < k^* + 1 \\ 1/(k^* + 1), & k \geq k^* + 1 \end{cases} \tag{2.35}$$

$$\eta = \min\{\gamma, 1 - \gamma\} \tag{2.36}$$

where $k^* = 1/\alpha$.

To summarize, when k is small ($k < k^*$), the system converges back to the same asymptotic behavior as $k = 0$ (Fig. 2.2cdg). In this region, k is not large enough to retain a good version once it appears. As a result, while performance might improve slightly in the first few attempts, it quickly saturates. In this region, agents reject prior attempts and thrash around for new versions, not processing enough feedback to initiate a pattern of intelligent improvement, prompting us to call it the *stagnation* region. Once k passes the critical threshold k^*, however, scaling behavior emerges (Fig. 2.2ceh), indicating that the system enters a region of *progression*, where failures lead to continuous improvement in both quality and efficiency. Nevertheless, with a single additional experience considered, the system quickly hits the second critical point k^*+1, beyond which the scaling exponent γ becomes independent of k (Fig. 2.2cfi). This means that once $\lceil k^* \rceil + 1$ number of prior failures are considered, the system is characterized by the same dynamical behavior as $k \to \infty$, indicating that $\lceil k^* \rceil + 1$ attempts are sufficient to recover the same rate of improvement as considering every failure from the past.

2.2.5. Phase transitions

Importantly, the two critical points in our model can be mapped to phase transitions within a canonical ensemble consisting of three energy levels (Fig. 2.3g-j). Phase transitions indicate that small variations at the microscopic level may lead to fundamentally different macroscopic behaviors. For example, two individuals near the critical point may initially appear identical in their learning strategy or other characteristics, yet depending on which region they inhabit, their outcomes following failures could differ dramatically (Figs. 2.2jk). In the progression region ($k > k^*$), agents exploit rapid refinements to improve through past feedback. By contrast, those in the stagnation region ($k < k^*$) do not seem to profit from failure, as their efforts stall in efficiency and saturate in quality.

To understand the nature of two transition points of our model, here we consider a canonical ensemble of N particles ($N \to \infty$) and three energy states $E_a(h) = 1$, $E_b(h) = (2h-1)^2$, and $E_c(h) = 1$, where h denotes the external field. We can write down the partition function of the system $Z = e^{-NE_a(h)} + e^{-NE_b(h)} + e^{-NE_c(h)}$, and calculate its free energy density $f = \ln Z/N$. In this system, it can be shown that the magnetization density $m = \frac{df}{dh}$ is discontinuous at the boundary of two energy states $E_a(h) = E_b(h)$ and $E_b(h) = E_c(h)$, characterized by two phase transitions at $h = 0$ and $h = 1$, respectively.

We notice that the canonical ensemble considered above has a mapping to our model. Indeed, denoting with $\Gamma \equiv k^*\gamma/(1-\gamma)$ and $K \equiv k - k^*$, we can rescale the system as $\Gamma = \min\{\max\{\Gamma_a(K), \Gamma_b(K)\}, \Gamma_c(K)\}$, where $\Gamma_a(K) = 0$, $\Gamma_b(K) = K$, and $\Gamma_c(K) = 1$, allowing us to map the two systems through $f \to (2\Gamma - 1)^2$, $N \to \ln n$, $h \to K$, and $E_i(h) = [2\Gamma_i^2(K) - 1]^2$ (Fig. 2.3g-j).

To understand the origin of the two transition points, we can calculate the expected life span of a high-quality version, obtaining $\langle u(x) \rangle \sim \langle (1-x)^{-\min\{k/k^*, 1/k^*+1\}} \rangle$. The first critical point k^* occurs when the first moment $\langle u \rangle$ diverges. Indeed, when k is small ($k < k^*$), $\langle u \rangle$ is finite, indicating that high-quality versions can only be reused for a limited period. Once k passes the critical point k^*, however, $\langle u \rangle$ diverges, offering the possibility for a high-quality version to be retained for an unlimited period of time. The second critical point arises due to the competition between two dynamical forces: (i) whether the current best version becomes forgotten after k consecutive attempts in creating new versions (dominated by the k/k^* term); or (ii) it is substituted by an even better version (dominated by the $1/k^* + 1$ term).

Note that while phase transitions carry exceptional importance in statistical physics, similar phenomena and concepts are also of fundamental relevance in the social/behavioral science literature. For example, critical thresholds have been observed and modeled in social settings ranging from shifts in neighborhood segregation [45] to social network formation [46] to collective opinion change [47]. In each case, slight shifts in micro-scale phenomena, like average preference, group size, or interaction intensity, condition a qualitative transition in macro-scale outcomes.

2.2.6. Modeling failure streak length

To understand the fat-tailed distribution documented in Fig. 2.1, let us consider a single-component case of our model for simplicity. We assume that q, the probability for a new version to success, is independent of its score. We denote N as the number of failures before success.

Assume $N \geq n$, i.e. one has not achieved success in the first n attempts. For one to succeed in the $(n+1)$-th attempt, she needs to (1) create a new version at this time, corresponding to probability $t_n \sim n^{-\gamma}$ and (2) succeed for this new version, which has probability q. Together we obtain

$$P(N = n | N \geq n) \sim q n^{-\gamma} \tag{2.37}$$

Note that this form is closely related with Lindy's law [48, 49]. Here the right hand side of the equation is decreasing, since a long failure streak indicates the existence of an (unsuccessful) version that has been used for a long period. Therefore, the same version is more likely reused again in the future, reducing the chance to create a new, successful version at the next step.

If we define the survival function $S(n) = P(N \geq n)$, this equation is equivalent to

$$1 - S(n+1)/S(n) \sim q n^{-\gamma} \tag{2.38}$$

Using a continuous approximation we obtain

$$-\frac{dS}{S} \sim qn^{-\gamma}dn \tag{2.39}$$

leading to the solution

$$P(N \geq n) = S(n) \sim e^{-cn^{1-\gamma}} \tag{2.40}$$

Hence, it predicts that the length distribution follows the well-known Weibull distribution.

To further understand the Weibull form, here we recognize that it is closely related to Heaps' law [50] caused by the reuse mechanism. Indeed, given that one needs to create M different versions before success, the distribution can be formulated as an exponential model

$$P(M \geq m) = (1-q)^m \tag{2.41}$$

However, repeated reuse leads to a sub-linear scaling between N and M, following the Heaps' law with exponent $1-\gamma$:

$$M(N) = \sum_{n=1}^{N} t_n \sim \sum_{n=1}^{N} n^{-\gamma} \sim N^{1-\gamma} \tag{2.42}$$

Combining the two equations one can obtain the same Weibull model

$$P(N \geq n) = S(n) \sim e^{-cn^{1-\gamma}} \tag{2.43}$$

2.3. Empirical methods

2.3.1. Length distribution of failure streaks

The length distribution of the failure streak is measured directly from data and fitted using maximum likelihood estimation techniques [51]. We fit empirical data with discrete version of Weibull (stretched exponential) form using maximum likelihood estimation with parameters $x_{\min} = 2$ and calculate uncertainty from bootstrapping over 100 simulations, yielding $\beta_1 = 0.666 \pm 0.017$, $\beta_2 = 0.566 \pm 0.086$, and $\beta_3 = 0.129 \pm 0.033$. Comparing this with γ estimated from temporal dynamics, two-sided t-tests indicate that none of the three datasets can reject the validity of the scaling identity $\beta + \gamma = 1$ ($P = 0.176,\ 0.421,\ 0.141$). We further compare the fitting results from alternative models, i.e. lognormal, power law, and truncated power law using likelihood ratio test [51], finding that Weibull distribution is consistently among the best functional forms (Table 2.1). To quantify the uncertainty of parameter estimations, we performed bootstrapping technique (100 times) to calculate optimal estimation for each round, and obtained standard error of parameter estimators. We also repeated the results for $x_{\min} = 3$, obtaining $\beta_1 = 0.592 \pm 0.032$, $\beta_2 = 0.513 \pm 0.175$, and $\beta_3 = 0.139 \pm 0.060$, which again statistically supports $\beta + \gamma = 1$.

2.3.2. Measuring failure dynamics

Given the highly skewed distributions of N and t_n, to measure $T_n = t_n/t_1$ we first performed log transformation to calculate the mean and variance of $\log(T_n)$ from

$$\mathrm{E}[\log(T_n)] = \langle \log(t_n/t_1) \rangle \tag{2.44}$$

	Exponential	Lognormal	Power law	Truncated power law
NIH grants	0.0	0.154	7.01×10^{-4}	2.33×10^{-159}
Startups	7.01×10^{-5}	0.723	2.48×10^{-6}	0.953
Terrorist attacks	0.0	0.822	0.566	0.221

Table 2.1. **Comparing different functional forms of distributions with Weibull distributions.**

All P-values terms denote the degree that Weibull distribution is compared over the other in log-likelihood ratio tests. Among all alternatives, only lognormal models show comparable fitting performance. Yet lognormal model uses two free parameters while the shape parameter of Weibull distribution is constrained by the scaling identity.

$$\mathrm{Var}[\log(T_n)] = \langle[\log(t_n/t_1)]^2\rangle - \langle\log(t_n/t_1)\rangle^2 \tag{2.45}$$

where we take $t_n = \max\{t_n, 1\}$ when necessary. We have also checked the robustness of this operation by trying to replace 1 with 0.5, finding the results remain similar. As the number of samples decreases dramatically with n, here we focus on $n \leq 10$ for D_1, $n \leq 7$ for D_2, and $n \leq 4$ for D_3.

The two equations immediately give us mean $\mathrm{E}[\log(T_n)]$ and standard error of the mean $\sqrt{\mathrm{Var}[\log(T_n)]/\text{sample size}}$, as plotted in Fig. 2.4. The divergence between the two groups can be detected as early as the second attempt. Although $T_1 \equiv 1$ by construction, Student's t-test rejects the hypothesis that $\log(T_2)$ between success and unsuccessful groups are the same ($P = 0.000457$, 0.00773, and 0.0992, respectively).

To calculate the temporal scaling exponent γ, here we run linear regressions between $\log(n)$ and $\log(T_n)$ and take the negative slope as γ, i.e.

$$\log(t_n/t_1) = -\gamma\log(n) + c, \tag{2.46}$$

yielding $\gamma_1 = 0.361 \pm 0.010$, $\gamma_2 = 0.509 \pm 0.036$ and $\gamma_3 = 0.640 \pm 0.153$ for successful group, with $P < 0.001$ for all three datasets. We also performed individual fixed effect linear models using samples with at least three data points, i.e.

$$\log(t_{n,j}) = -\gamma \log(n) + c_j + \epsilon_{n,j}, \tag{2.47}$$

where j is the index for different samples and c_j is the fixed effect term for each agent j. We obtain similar results $\gamma_1 = 0.372 \pm 0.017$, $\gamma_2 = 0.431 \pm 0.077$ and $\gamma_3 = 0.685 \pm 0.182$. For unsuccessful group there exists no significant relationships between $\log(n)$ and $\log(T_n)$ since the second failure (i.e. excluding T_1), with $P = 0.450$, $P = 0.884$ and $P = 0.957$ respectively. Together, these results offer strong empirical support for the diverging temporal patterns predicted by our model.

2.4. Testing model predictions

As such, the phase transitions uncovered in our simple model make four distinct predictions, which we now test directly in the contexts of science, entrepreneurship, and security.

2.4.1. Prediction A

Not all failures lead to success

While we tend to focus on examples that eventually succeeded following failures, the stagnation region predicts that there exists a non-negligible fraction of cases that do not succeed following failures. We measure the number of failed cases that did not achieve eventual success in our three datasets, finding that members of the "non-success" group not

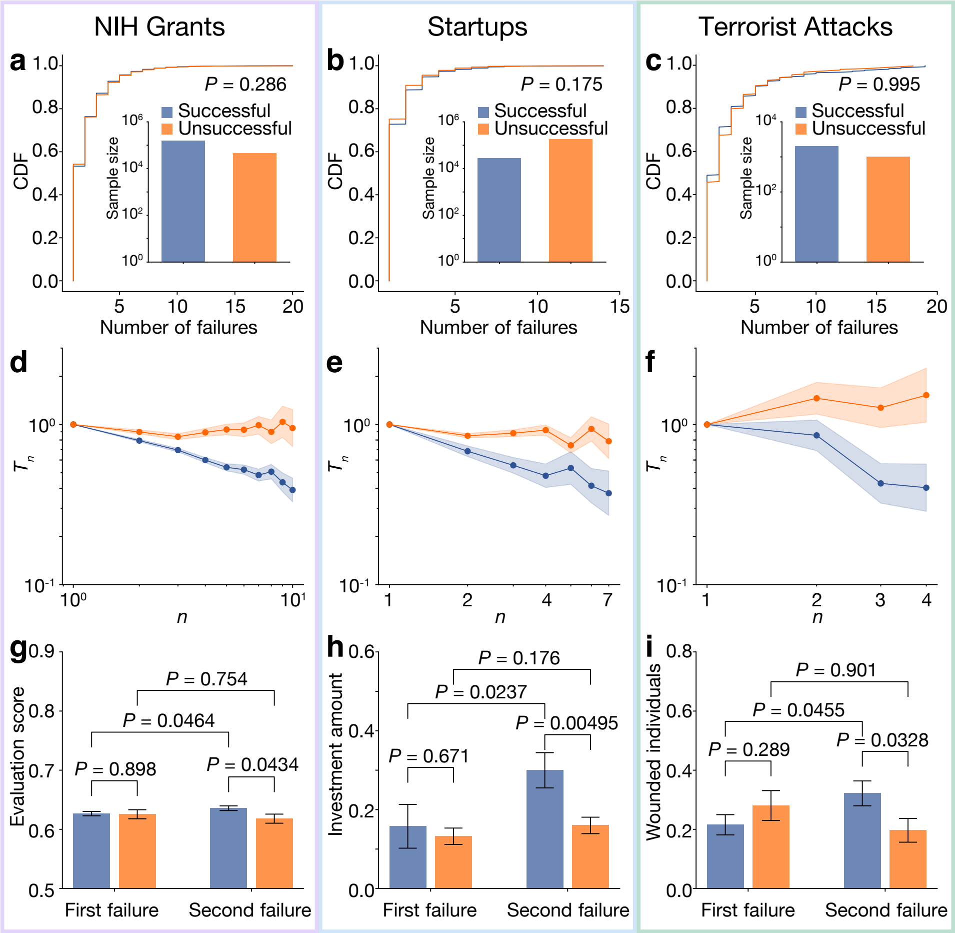

Figure 2.4. **Testing model predictions.**

(a-c) Cumulative distribution function (CDF) of the number of consecutive failures prior to the last attempt for the success and non-success groups. To eliminate the possibility that agents were simply in the process of formulating their next attempt, we focus on cases where it has been at least five years since their last failure. In each of our three datasets, two distributions are statistically indistinguishable. For clarity, here we show results for less than 21 failures. (Inset) The sample size of success and non-success group, showing their size is of a similar order of magnitude. **(d-f)** Early temporal signals separate success and non-success groups. For each group we measure the average inter-event time between two failures $T_n \equiv t_n/t_1$ as a function of the number of attempts. Dots and shaded areas show the mean and s.e.m. measured from data. All success groups manifest power law scaling $T_n \sim n^{-\gamma}$. The two groups show distinguishable temporal dynamics for $n = 2$. This temporal scaling is absent for non-success groups. (cont. on next page)

(Previous page.) **(g-i)** Performance at first attempt appears indistinguishable between the success and non-success groups who experienced a large number of consecutive failures prior to the last attempt, but becomes distinguishable from the second attempt. Whereas performance improves for the success group, this improvement is absent for the non-success group. The center and error bar show the mean and s.e.m.

only exist, but their size is of similar order of magnitude as the success group (Figs. 2.4a-c inset). Interestingly, the number of consecutive failures prior to the last attempt for the non-success group follows a statistically indistinguishable distribution from those that lead to success (Figs. 2.4a-c), suggesting that people who ultimately succeeded did not try more or less than their non-successful counterparts.

2.4.2. Prediction B

Early dynamical signals separate the success group from the non-success group

The model predicts that the success group is characterized by power-law temporal scaling, which is absent for the non-success group (Fig. 2.2j), predicting the two groups may follow fundamentally different failure dynamics distinguishable at an early stage. To test this prediction, we measure the average inter-event time between two failures T_n as a function of the number of failures. Figures 2.2d-f unveil three important observations. (i) For the success group, T_n decays with n across all three domains, approximately following a power law, as captured by Eq (2.12). The scaling exponents are within a similar range as those reported in learning curves [27], further supporting the validity of power law scaling. Although the three datasets are among the largest in their respective domains, agents with a large number of failures are exceedingly rare, limiting the range of n that can be measured empirically. We therefore test if alternative functions may offer a better fit, finding power law to be the consistently preferred choice. (ii) Temporal scaling disappears, however, when we measure the same quantity for the non-success group (Figs. 2.4d-f), consistent with predictions about the stagnation region. (iii) The two groups show distinguishable failure dynamics as early as $n = 2$, suggesting intriguing early signals that separate those who eventually succeed from those who do not.

Observations uncovered in Figs. 2.4d-f are intriguing for two main reasons. First, failures captured by the three datasets differ widely in their scope, scale, definition, and temporal resolution, yet despite these differences, they are characterized by remarkably similar dynamical patterns predicted by our simple model. Second, while one might expect that the last attempt was crucial in separating the two groups, as the model predicts, success and non-success groups each follow their respective, highly predictable patterns, distinguishable long before the eventual outcome becomes apparent. We use a simple logistic model to predict whether one may achieve success following N previously failed attempts in D_1, using only temporal features t_n $(1 \leq n \leq N-1)$ as predictors. To evaluate prediction accuracy, we calculate the AUC curve over 10-fold cross validation. We find that, by observing timing of the first three failures alone, our simple temporal feature yields high accuracy in predicting the eventual outcome with an AUC close to 0.7, significantly higher than random guessing (Mann-Whitney rank test, $P < 10^{-180}$, Fig. 2.5a). We repeated the same prediction task on D_2 and D_3, arriving at similar conclusions (Fig. 2.5b,c). The predictive power from temporal features alone is somewhat unexpected. Indeed, there are a large number of documented factors that affect the outcome of a grant application [7, 52–55], ranging from prior success rate to publication and citation records to race and ethnicity of the applicant. Yet here we ignore these

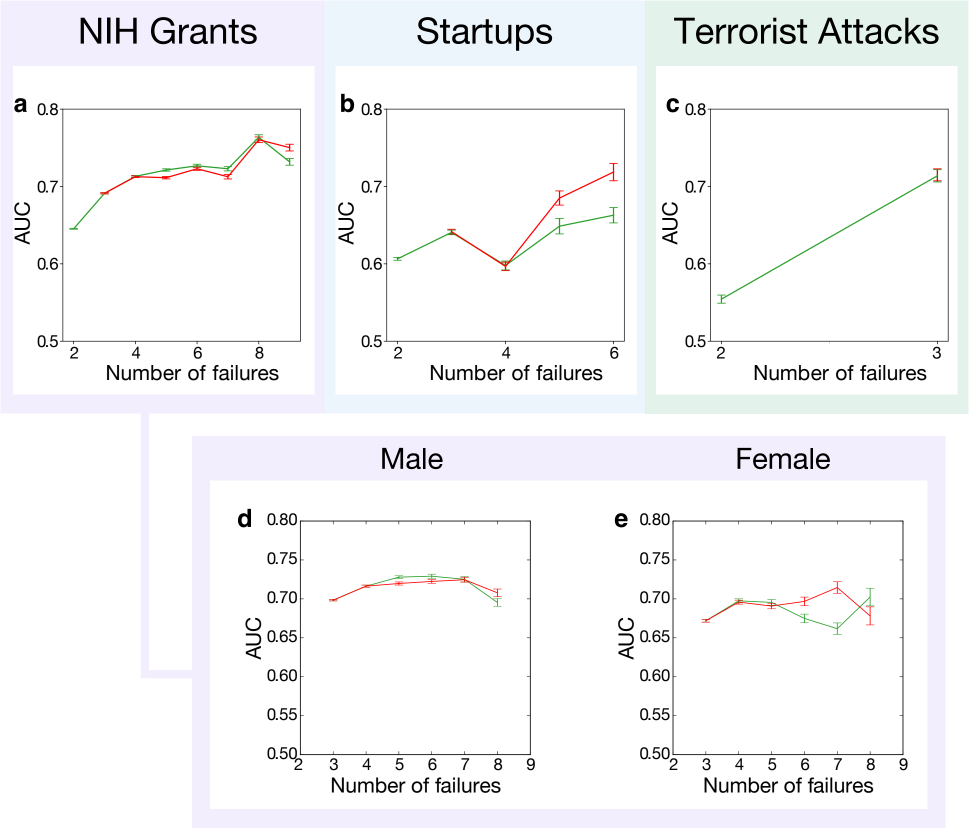

Figure 2.5. **Predicting ultimate success in science, startups and security.**

a-c, Area Under the curve of the Receiver Operating Characteristic (AUROC) of the prediction task. We apply two logistic regression models to predict ultimate success in NIH grants (**a**), startups (**b**) and terrorist attacks (**c**). The centers and error bars of AUROC scores denote the means and s.e.m. calculated from 10-fold cross validation over 50 randomized iterations (green: Model 1, red: Model 2). **d-e**, As in **a** but predicting ultimate success in NIH grants for male (**d**) and female (**e**) investigators.

factors, using only features pertaining to temporal scaling as prescribed by our model. This suggests that our predictive power represents a lower-bound, which could be further improved and leveraged by incorporating additional factors.

To test if the observed patterns in Figs. 2.4d-f may simply reflect preexisting population differences, we take agents who experienced a large number of failures, and measure performance from their first attempt. We find that for all three domains, the two populations were statistically indistinguishable in their initial performance (Figs. 2.4g-i), which leads us to the next prediction:

2.4.3. Prediction C

Diverging patterns of performance improvement

Although the two groups may have begun with similar performance, the model predicts they may experience different performance gains through failures (Fig. 2.2k). We compared performance at first and second attempts, finding significant improvement for the success group (Figs. 2.4g-i), which is absent for the non-success group.

One key difference between progression and stagnation regimes is the propensity to reuse past components. From the perspective of exploration vs. exploitation [56, 57], however, reuse helps one retain a good version when it appears, but it could also keep one in a suboptimal position for longer, suggesting our final prediction:

2.4.4. Prediction D

The length of failure streaks follows a Weibull distribution

$$P(N \geq n) \sim e^{-(n/\lambda)^{\beta}}. \tag{2.48}$$

	NIH grants	Startups	Terrorist attacks
γ	0.361 ± 0.010	0.509 ± 0.036	0.640 ± 0.153
β	0.666 ± 0.017	0.566 ± 0.086	0.129 ± 0.033
P	0.176	0.421	0.141

Table 2.2. **Parameter estimates (mean$_{\pm\mathbf{s.e.m}}$).**

γ corresponds to the temporal scaling exponent uncovered in Fig. 2.4 d-f) and β is the shape parameter of the Weibull distribution (s.e.m. estimated from bootstrapping over 100 simulations), characterizing the length distribution of failure streaks. Two-sided t-tests indicate that none of the three datasets can reject the validity of the scaling identity $\beta + \gamma = 1$.

Moreover, the shape parameter β is connected with the temporal scaling exponent γ through a scaling identity

$$\beta + \gamma = 1. \tag{2.49}$$

This means that if we fit the streak length distribution in Figs. 1h-j to obtain the shape parameter β, it should relate to the temporal scaling exponent γ, obtained from Figs. 2.23d-f. Comparing β and γ measured independently across all three datasets shows consistency between our data and the scaling identity (Table 2.2).

2.5. Model extensions

As a single parameter, k necessarily combines individual, organizational and environmental factors in learning [10, 37]. The one-parameter model developed here represents a minimal model, which can be extended into richer frameworks. For example, agents may have varied incentives to improve or may differ in their confidence and ability to judge their prior work. Such factors trace heterogeneity in the population and can be captured by the α parameter, which quantifies individuals' propensity to change given feedback. This leads us to develop the $k-\alpha$ model, which predicts a two-dimensional phase diagram with three distinct phases. The model can be further extended to capture fuzzy inference from past feedback, allowing agents to not always choose the best prior versions.

2.5.1. $k - \alpha$ model

Agents may differ in the judgment of their own work or incentives to change given feedback, which can be captured by varying the α parameter in the original k-model. Of the many influences on p, one key factor is the quality of existing versions, suggesting that p should be a function of x^*. Consider the two extreme cases: If $x^* \to 0$, existing versions of this component have among the worst scores and, hence, a high potential for improvement when replaced with a new version. Indeed, the likelihood of creating a new version is high, i.e., $p \to 1$. On the other hand, $x^* \to 1$ corresponds to a near-perfect version, yielding a decreased incentive to create a new one ($p \to 0$). Indeed, $P(x \geq x^*) = 1 - x^*$ captures the potential to improve on prior versions, prompting us to assume $p = (1-x^*)^\alpha$, where $\alpha > 0$ characterizes an agent's propensity to create new versions given the quality of existing ones. Therefore, $\alpha \to 0$ indicates that regardless of one's evaluation, the agent

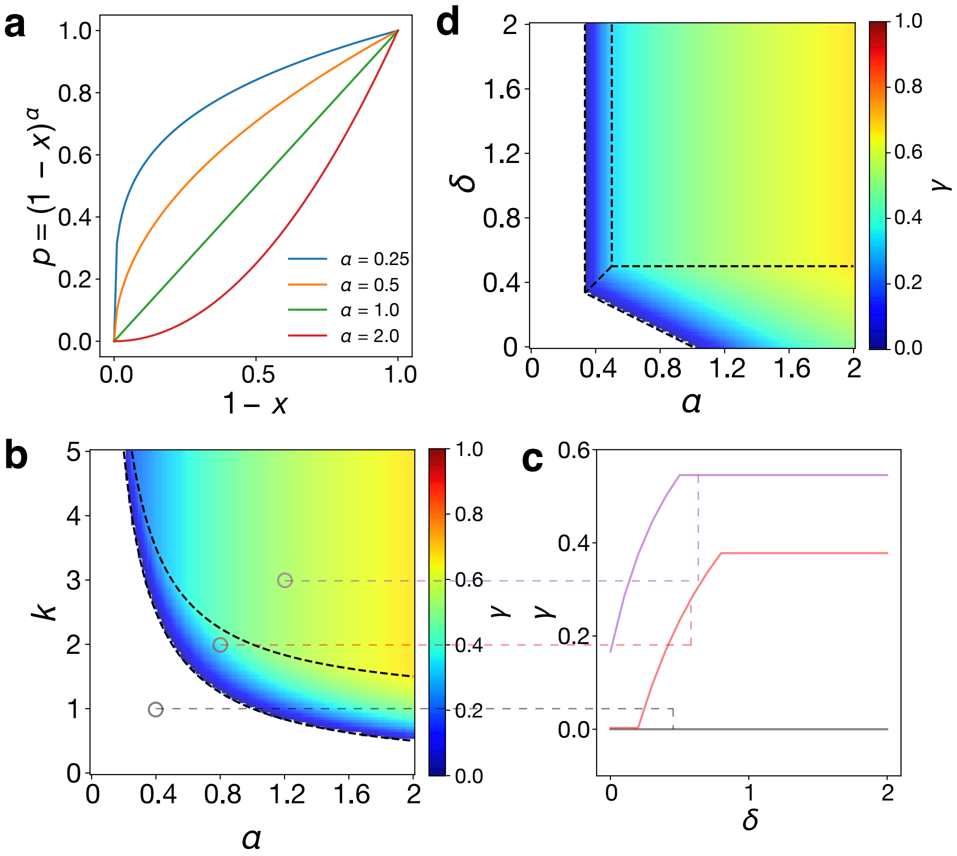

Figure 2.6. **Generalization of the k model.**

a, The α parameter connects the potential to improve $1 - x$ and likelihood to create new versions p through $p = (1 - x)^{\alpha}$. **b**, Phase diagram of the $k - \alpha$ model. The two-dimensional parameter space is separated into three regimes, with boundaries at $k\alpha = 1$ and $(k - 1)\alpha = 1$. **c**, The impact of δ parameter on scaling exponent γ for given of $k = 1, 2, 3$ and $\alpha = 0.4, 0.8, 1.2$. We find δ affects the temporal scaling parameter when it is small, but has no further impact beyond a certain point $\delta^{*} = \min(\alpha, 1/k)$. **d**, Phase diagram of the $k - \alpha - \delta$ model for $k = 3$, with boundaries at $\alpha = \delta$, $(k - 1)\delta = 1$, $(k - 1)\delta + \alpha = 1$, $k\alpha = 1$, and $(k - 1)\alpha = 1$, respectively.

will always create a new version, whereas $\alpha \to \infty$ points to the other extreme where one does not create a new version unless it is extremely bad (Fig. 2.6a). Considering α as another tunable parameter, we arrive at a two-parameter model: the $k - \alpha$ model .

To solve this model we can substitute k^* with $1/\alpha$, and the indexes k/k^* and $1/k^*+1$ now become $k\alpha$ and $\alpha+1$. The extended model thus predicts the existence of three different phases on a two-dimensional phase diagram, with the boundaries $k\alpha = 1$ and $(k-1)\alpha = 1$ that separate the three phases (Fig. 2.6b). The $k-\alpha$ model reduces back to the two critical points in the original k model when we fix α. The two parameters jointly define an 'effective' $K \equiv k - k^* = k - 1/\alpha$. The critical boundaries therefore reduce into two simple equations: $K = 0$ and $K = 1$. Note that the assumed relationship between p and $(1-x^*)$ is not limited to a power law but can be relaxed into its asymptotic form. Indeed, we show that as long as the function satisfies $\frac{\ln p}{\ln(1-x^*)} \to \alpha$ as $x^* \to 1$, the model offers the same predictions [41].

2.5.2. $k-\alpha-\delta$ model

Agents may have fuzzy inference of past feedback, hence may not always choose the version with the highest quality. We can model the choice between different versions in a probabilistic fashion, by introducing a δ parameter to the $k-\alpha$ model. Here the probability to choose the i-th version as a baseline follows

$$P(i) = \frac{1}{Z}(1-x_i)^{-\delta} 1_{n-k \le i \le n-1},$$

where Z is the normalization factor, $Z \equiv \sum_{i=n-k}^{n-1}(1-x_i)^{-\delta}$. $\delta = 0$ means one cannot differentiate quality between past versions and selects randomly among different versions, whereas $\delta \to \infty$ indicates that one always chooses the prior version with the highest quality, converging back to our original k model or the $k-\alpha$ model. Incorporating δ leads to the $k-\alpha-\delta$ model (Fig. 2.6cd).

Analytically solving the model reveals interesting scaling behaviors based on δ. Indeed, we find the scaling behavior of the system follows

$$\gamma(k, \alpha, \delta) = 1 - \{\max[\min(\alpha + (k-1)\min\{1, \alpha, \delta\}, \alpha + 1), 1]\}^{-1},$$

revealing rich mathematical properties. When $\delta \to \infty$, the new solutions converge back to the original solution for the $k - \alpha$ model. With δ the three-parameter model is characterized by four different phases. Three of the regimes are generalizations of those found in the $k - \alpha$ model, where the scaling exponent γ does not depend on δ in the limit of $\delta \to \infty$, i.e., $\gamma(k, \alpha, \delta) = \gamma(k, \alpha, \infty)$. The fourth one, however, is a new phase and only exists for small δ. The intuition is that, in this regime, the inability to select a high-quality version (small δ) dominates the scaling behavior, with exponent $\gamma(k, \alpha, \delta) = 1 - [(k-1)\delta + \alpha]^{-1}$. Together, these extensions offer further support for the predictions of our original model, while demonstrating the model's theoretical potential by enriching its mathematical properties with more realistic interpretations. They also point to promising future research that explores the interplay between different perspectives of learning.

2.6. Additional empirical observations

2.6.1. Quantifying component dynamics

In our modeling attempts, we treat components as purely abstract properties of a grant proposal, fledgling company, or terrorist campaign. Here we further consider if we can empirically measure or approximate components, thereby better estimating and understanding their dynamics and validating the descriptive power of our model. The difficulty of this measurement stems from the fact that the existing datasets obtained above, while extensive, are nevertheless inadequate in this respect. Indeed, unlike scientific papers, which have reference information that can approximate the units of knowledge they piece together, grant proposals are largely isolated documents, making it difficult to infer the ‘substance’ of each proposal. Furthermore, while some metadata are associated with each proposal, such as funding institute and PI affiliation, these data are typically constant for each individual applicant and hence useless for evaluating the dynamics of components across different attempts by the same individual.

To tackle these challenges, we acquired a new data corpus from the NIH that contains abstract information for all R01 proposals submitted after 2008 (both funded and unfunded). Since the abstract data is only available after 2008, and the definition of the unsuccessful group requires five years of inactivity, so there’s not enough data for us to measure the unsuccessful group. Nevertheless, the new data does offer a possibility for us to empirically measure the component dynamics for the successful group.

Our hypothesis here is that if we can perform content analysis on abstracts, it may allow us to measure components embedded in each new attempts. To achieve this, we applied a natural language processing (NLP) technique to NIH abstracts that estimated MeSH (Medical Subject Headings) terms associated with each proposal. Note that MeSH terms are one of the most commonly used classification codes for biomedical research [58], and this operation is only possible thanks to recent advances in NLP classification, allowing us to automatically and accurately infer MeSH terms from abstract texts. Specifically, we applied NLM Medical Text Indexer, an official protocol developed by US National Library of Medicine Indexing Initiative, to extract a list of MeSH terms given abstract texts.

While the obtained MeSH terms are necessarily imperfect and may not directly correspond to distinct components of the proposal, they capture information that reflects different facets of the proposal, including methods and experimental techniques (e.g., genomic screens), objects of analysis (e.g., breast cancer), research design (e.g., genome-wide association study), and physical phenomena (e.g., estradiol). Here we approximate the creation of new versions by the number of new MeSH terms (terms that did not appear in the previous k submissions), defined as m_n. For example, to measure the dynamics under $k = 1$, we count m_n as the number of Mesh terms that appear in the n-th attempt but not in the $(n-1)$-th attempt. More generally, if we define S_n as the set of all Mesh terms associated with the n-th attempt, our definition can be formulated as $m_n \equiv |S_n - (S_{n-1} \cup \cdots \cup S_{n-k})|$, where $|A|$ denotes the size of a set A (Fig. 2.7a).

According to our model, the time cost comes from creating new versions, traced by the introduction of additional components. Hence, our model suggests that given k, we can use $M_n \equiv \langle m_n \rangle / \langle m_1 \rangle$ to mimic the temporal dynamics of $T_n \equiv \langle t_n \rangle / \langle t_1 \rangle$. More precisely, for the successful group, we should expect to observe that for large k ($k > k^*$), M_n and T_n should be similar. Yet for small k ($k < k^*$), the two quantities should be quite different. This means that in the same way faster resubmissions (T_n) predict ultimate success, so do shrinking sets of new components (M_n).

We set out to test this new prediction by calculating M_n for different k. We find that the two curves follow different dynamics ($k \leq 3$). Yet the dynamics of M_n and T_n cannot be statistically distinguished for $k > 3$ (from 4 to ∞), both following a power law with $\gamma \sim 0.35$ (Fig. 2.7b). Both findings appear consistent with model predictions. Given that Mesh terms are merely a rough estimate of idea combination in NIH proposals, this degree of agreement seems unexpected.

2.6.2. Learning by organizational vs. individual

One aspect of our paper is that here we study learning processes at three different levels, ranging from individual attempts (PIs) to individuals in teams (entrepreneurs) to larger-scale organizations (terrorist groups). The patterns we uncovered reveal that all three levels follow similar statistical patterns governing failure dynamics. But beyond the universality, what differences should we expect across different levels? To answer this question, we contextualize our paper in the literature it builds upon.

The organizational learning literature has identified several factors for the emergence of learning within organizations, with some arguing that individual learning is just one factor in how and why organizations may learn. For example, knowledge gained from past experience can be embedded within both individual habits and organizational routines (including the idea of transactive memory) [24, 59, 60]. These suggest that organization-level learning, compared with individual learning, should be characterized by higher learning rate on average. There is also evidence that organizational learning tends to be conservative due to inflexible routines. For example, given versions with the same quality, organizations may have higher probability to reuse rather than create a new one.

Together, these theories predict that of the three domains studied, those closer to organizational learning (such as terrorists) should correspondingly have higher learning rates than those closer to individual learning (such as NIH PIs). We can test this hypothesis by calculating the average learning rate for our samples. We find that our estimations

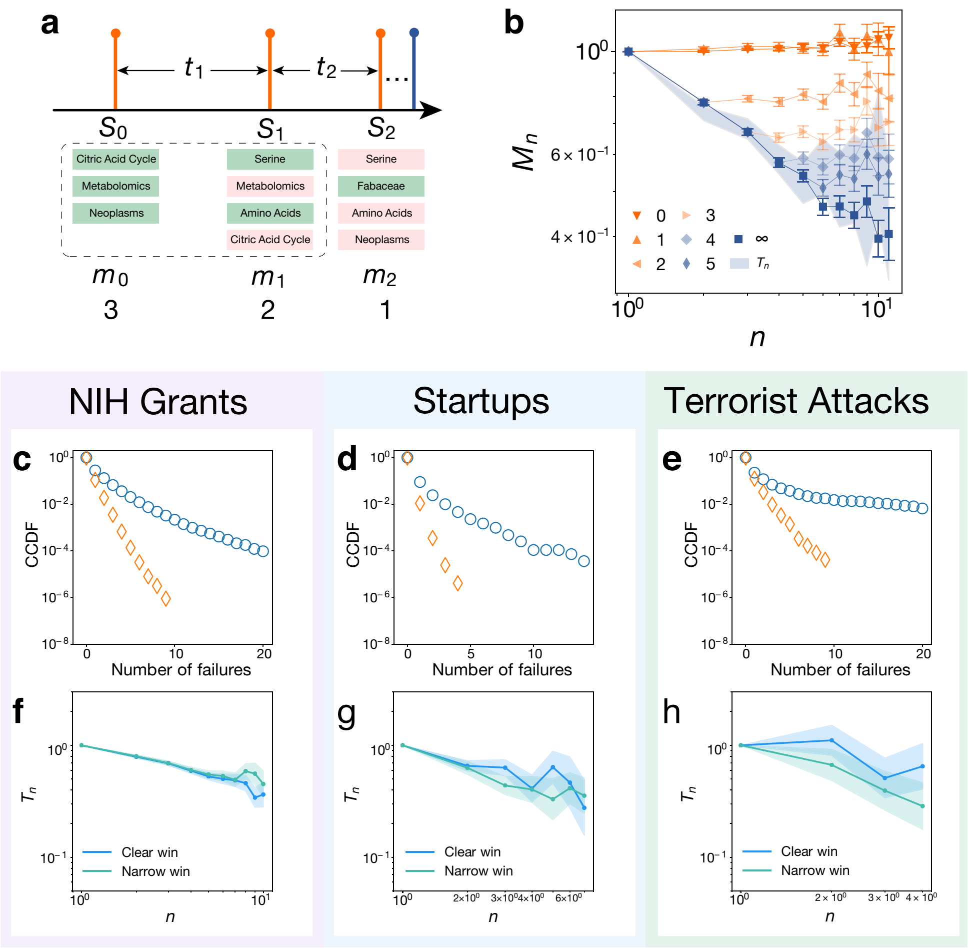

Figure 2.7. **Model validations.**

a, An illustration for component dynamics. We extract all MeSH terms associated with the n-th attempt, S_n, and calculate the number of new terms m_n, defined as $|S_n - (S_{n-1} \cup \cdots \cup S_{n-k})|$. **b**, Testing component dynamics in NIH grant applications. We calculate the dynamics of $M_n = \langle m_n \rangle / \langle m_1 \rangle$ using different k and compare it with T_n. The centers and error bars of M_n show the means and s.e.m. for different k. The shaded area shows mean ± s.e.m. of T_n measured on the same subset. All $k > 3$ lead to similar trends between M_n and T_n. **c–e**, Length of failure streak after randomization in science (c), entrepreneurship (d) and security (e). We take the samples used in Fig. 2.1 and shuffle the success/failure label from each attempt. This operation keeps both the overall success rate and the total number of attempts for each individual constant. **f–h**, Temporal scaling patterns within the successful group in science (f), entrepreneurship (g) and security (h). We separated the successful group into two subgroups (narrow winners and clear winners) based on eventual performance (0.9 in evaluation score for D_1, 0.5 in investment amount for D_2 and 1 in wounded individuals for D_3).

appear consistent with the hypotheses outlined above: For NIH PIs, the average learning rate γ is around ~ 0.361; The learning rate for the entrepreneurship case is higher, around ~ 0.509, and terrorist groups have the highest rate on average ~ 0.640. While these differences could be due to inherent domain-specific differences, they do show consistency with the theories from the organizational learning literature.

2.6.3. Scientific achievements and learning rate

Existing literature has also highlighted a series of factors related to why one learns more than others [61], including individual ability, motivation and opportunity to learn. These factors may play a role, manifested in the k parameter. One empirical challenge here is that it remains unclear how to infer k directly from data. But we also realize we can relax the assumption to infer a weak form of the parameters by inferring γ, and correlate individual characteristics (y) with γ. Indeed, according to our model, if y correlates with k, it may not correlate with γ (if it's in the third phase ($k > k^* + 1$)), but if y correlates with γ, then it must correlate with k.

High achieving scientists are more visible, better recognized, and have access to more resources (Matthew effect in science) [62–64], suggesting that individual prior achievement may manifest in a higher learning rate [61]. Here we test this hypothesis from our data, by collecting additional datasets that allow us to identify individual characteristics and achievements.

Here we extend our analysis of individual characteristics by linking NIH data to the Web of Science citation database. This procedure involved systematic effort in paper matching and author name disambiguation. Here we began from a list of NIH supported publications in PubMed and selected those authored by the same PI. Then we use a WoS-PubMed crosswalk file to locate these papers in WoS and treated them as ‘seed’ papers. We then expand this initial set to other publications by the same-name author in WoS by tracing the citation relationships and following standard name disambiguation procedures [17, 65]: If a paper was contributed by an author with the same name and had citation/reference/co-reference relationships with the initial set, we included it into the PI publication list as well. Implementing this method iteratively allowed us to construct a comprehensive publication list for each PI in our sample.

We then calculated the learning rate γ by regression for all samples with at least three failures before eventual success (i.e., more than two inter-event time periods). Based on the learning literature, we hypothesize that the learning rate may be related to experiences both within and outside the task of producing an NIH proposal [24]. To this end, we calculated the total number of citations of a PI for all his/her papers published before the first failure (logged), approximating his/her overall ‘status’ and accomplishments. We find that it is significantly, positively correlated with the learning rate γ ($P < 0.001$, after controlling for the first inter-event time). We further test this correlation by including the number of prior successes and application year as control variables, finding that although past funding success is also correlated with higher learning rate ($P < 0.001$), the relationship between citations and γ remains robust ($P = 0.014$). Although it may seem intuitive that citations and grant applications are correlated, note that the samples studied here include PIs who all failed at least three times before eventually being awarded the grant (i.e., similar success rate). In this respect, it is somewhat unexpected to observe that the speed with which scientists learn from failures can be anticipated by measuring prior achievements. This is consistent with the hypothesis that prior attention and success may provide scientists with greater confidence and resources that allow them to persist and refine rather than abandon and replace the components from an initial, failed proposal.

2.6.4. Gender and learning rate

The results presented above offer support for the notion that individual characteristics can indeed affect learning. Here we further anchor other individual characteristics that may distinguish learning. The literature suggests gender could be a potential robust factor that applies across domains, especially in science and entrepreneurship, which are characterized by persistent gender inequality [66–70]. It thus suggests that, if we can separate individuals by gender, we may detect differential learning rates as well.

To test this relationship in our data, we use a gender detector algorithm to infer gender information from person's first name. We find that gender indeed plays an important role, after we control for all other factors. Our regression analysis shows significant correlation between gender and learning rate. All other factors being equal, the learning rate γ of a male PI in NIH system exceeds that of a female PI by 0.14 ($P = 0.001$). That is, male PIs fail faster than their female colleagues. This difference appears substantial, considering that the average learning rate is centered around 0.35. Note that here we do not essentialize these gender differences, and recognize that they may flow from institutional as well as individual causes, such as a culture that discourages women from persistence and encourages oversensitivity to feedback. Furthermore, such correlations cannot fully account for the discovered signals, as a substantial amount of predictive power by our model remains (AUC higher than 0.7) after we separate our samples by gender.

We further test this relationship on startup dataset, finding a similar gap of 0.10 in the same direction between male and female innovators, though the result is not significant, possibly due to a smaller sample size. These insights are consistent with existing literature on gender inequality in science and entrepreneurship [66–69]. They also highlight the fact that our paper offers a new theoretical framework to systematically study learning, failures and the factors that may influence them.

2.7. Related works

2.7.1. Learning literature

This work is closely related to the rich literature on learning and failures. Canonical frameworks in understanding how people react to failures [12, 25, 71–75] have identified several key factors that could impact learning, including individual characteristics and organizational structures and strategies. These findings have also prompted quantitative studies using failure records across different industries, ranging from entrepreneurship [9, 10] to commercial banking [76], from healthcare [77] to coal mining [78] to trains [79], and airlines [80] to orbital launch vehicles [81].

Another relevant line of inquiry is in psychology and organization behavior, which concerns learning curves from both theoretical [24–27, 32, 37, 41, 75, 82–89] and empirical [24, 27, 39, 42, 75, 82, 90] perspectives, quantifying how performance and efficiency improve with experience. One key result is the famous Wright's law [42], i.e. the power law form of cost reduction.

Next we review a series of major models and compare key predictions with our empirical results. We summarize all these models in Table 2.3.

Category	Reference	Time	Performance	Power law	Coexistence
Adaptation	Crossman[91]	✓	✗	✗	✗
	Kauffman & Levin[92]	✗	✓	✗	✗
	Denrell & March[33]	✗	✓	✗	✗
Search	Roberts[86]	✗	✓	✓	✗
	Muth[41]	✗	✓	✓	✗
	Mcnerney *et al*[32]	✗	✓	✓	✗
Individual learning	Newell *et al*[39]	✓	✗	✓	✗
	Anderson[40]	✓	✗	✓	✗
Urn	Simon[93]	✗	✗	✓	✗
	Tria *et al*[94]	✓	✗	✓	✓
	Iacopini *et al*[95]	✓	✗	✓	✓
Other	Levy[82]	✗	✓	✗	✗
	Shrager *et al*[87]	✗	✓	✗	✗
	Sahal[85]	✓	✗	✓	✗
	Johnson *et al*[37]	✓	✗	✓	✗
	Clauset & Gleditsch[88]	✓	✗	✓	✗

Table 2.3. **Literature review of relevant models.**

We test whether the models listed can predict (1) Time: time reduction; (2) Performance: performance improvement (or reduction in any cost other than time); (3) Power law: analytical form of power law scaling; (4) Coexistence: coexistence of two groups with different dynamics (success and unsuccessful groups in this paper). We find that none of the existing models can predict all the observations in our paper.

2.7.2. Stochastic models with memory

One school of thought can be viewed as modeling the dependence structure among failures. Indeed, the failure of the chance model suggests that non-trivial dependence may be essential for modeling the fat-tailed length distribution of failure streaks, which raises an important question: Could other stochastic processes (Markov process, random walk, autoregressive model, etc.) account for our observations? Indeed, if we consider a general framework of fixed dependence as follows

$$S_n = f_n(S_1, S_2, \cdots, S_{n-1}), \tag{2.50}$$

where S_n denotes the performance at the n-th attempt and f_n can be a deterministic or stochastic non-decreasing mapping. This framework covers a wide range of stochastic processes, e.g. $f_n(S_1, \cdots, S_{n-1}) = f_n(S_{n-1})$ for a discrete space of S_n leads to Markov process, $f_n(S_1, \cdots, S_{n-1}) = S_{n-1} + \epsilon_n$ leads to random walk, $f_n(S_1, \cdots, S_{n-1}) = \sum_{i=1}^{p} \phi_i S_{n-i} + \epsilon_n$ leads to autoregressive model. We note that if this is true, we can obtain

$$S_n = f_n(S_1, f_1(S_1), \cdots, f_{n-1}(S_1, f_1(S_1), \cdots)) \equiv g_n(f_1, \cdots, f_n)(S_1) \tag{2.51}$$

Hence, S_n can be formulated as a non-decreasing function of S_1, indicating that there should be detectable 'fitness' differences in the first attempt. Indeed, these results indicate that if there exists no difference in the dependency structure f_n, the differences in outcomes should be at least partly contributed by performance at the first attempt, which contradicts with our data. This hypothesis also cannot explain the fat-tail length distribution of failure streaks (S3.8).

2.7.3. Adaptation models

The evolutionary perspective for individual and organizational learning assumes that the agent improves through updating information and belief on different alternatives. Here we discuss three representative models, each assuming a finite pool of available options.

2.7.3.1. Crossman's model. Crossman's model, first proposed in [91], aims to explain the temporal dynamics observed in individual tasks. The model suggests a process from r methods M_i ($1 \leq i \leq r$), each with a time cost t_i. The individual improves operation strategy through changing probabilities for using different methods, i.e. p_i where

$\sum_{i=1}^{r} p_i = 1$. At the n-th trial, the expected time cost can be formulated as

$$T(n) = \sum_{i=1}^{r} t_i p_i(n) \tag{2.52}$$

The change of probability for choosing method M_i is proportional to the difference between its time cost and current average time cost, i.e.

$$p_i(n+1) - p_i(n) = -k(t_i - T(n)) \tag{2.53}$$

Therefore, the time cost decays as

$$T(n+1) = T(n) - k \sum_{i=1}^{r} p_i (t_i - T(n))^2 \tag{2.54}$$

2.7.3.2. NK model. NK model, initially proposed by Kauffman [92] is a canonical model in organizational learning [96]. Consider a rugged fitness space of N dimensions $X = (x_1, \cdots, x_N)$, where $x_i \in \{0, 1\}$. The fitness score of each possibility is the summation of interaction among K adjacent dimensions, that writes

$$\phi(x) = \sum_{i=1}^{N} \phi_i(x_i, \cdots, x_{i+K}) \tag{2.55}$$

One heuristic searching strategy in this rugged landscape concerns two options:

(1) Local search, i.e., walk to a neighbor, y, which satisfies $|y - x| = 1$.

(2) Global search, i.e., jump to a new node randomly.

2.7.3.3. Denrell and March's model. Denrell and March proposed a simple adaptation model to understand the interplay between information and adaptation, explaining why people have bias against novel and risky choices [33]. In this model, P_t, defined as the probability for the first option to be chosen at time t, depends on its past probability P_{t-1} and current performance. If the option leads to better outcome compared with the other, one updates

$$P_{t+1} = P_t + a(1 - P_t) \tag{2.56}$$

otherwise,

$$P_{t+1} = (1 - a)P_t \tag{2.57}$$

All three models presented here can mimic specific performance or efficiency trajectory as one tries repeatedly. The main issue with these models is that they all base on a finite space of possible options, which leads to a limit in performance and efficiency improvement that one cannot overcome, which contradicts with our data.

2.7.4. Search models

Search models assume an iterative process, where one decides whether to use existing components or try new ones based on component quality. Such models are often characterized by an improvement in the objective performance function because of the extreme values theory, i.e. as one always selects the best version from experimentation, she will eventually arrive at the version that is reasonably good.

2.7.4.1. Roberts' model. Robert proposed a model based on greedy algorithms [86]. To understand the universal learning process, the model assumes production efficiency p

as lognormal, following

$$x = b \ln p \tag{2.58}$$

where x follows the standard normal distribution $N(0,1)$. Each time the agent randomly selects a sample x' and compares it with current efficiency x, adopting the new method when $x' < x - a$. The model predicts

$$\ln p \sim \ln N/ab \tag{2.59}$$

2.7.4.2. Muth's model. Muth's model [41] builds on a simple assumption: the individual tries a new method at each trial and uses the new method if it costs less. The model further assumes appropriate regularity conditions for the cumulative distribution function (CDF) of cost F, e.g.

$$\lim_{x \to x_0} \frac{F(x)}{(x - x_0)^k} = c \tag{2.60}$$

where x_0 is the limiting cost of production. The model predicts the expected cost $E[X_n]$ of the n-th production as

$$E[X_n] = x_0 + \Gamma(1 + 1/k)(cn)^{-1/k} \tag{2.61}$$

Muth's model is an elegant model explaining the emergence of power law scaling and can be extended to dependent component cases.

2.7.4.3. McNerney's model. McNerney et al further extended Muth's model by assuming a power law distribution of costs of each component ($f(c_i) \sim x_i^{\gamma-1}$) and using design structure matrix to characterize the dependency among different components [32]. The model predicts the cost y decreases as a function of productions n following

$$y(n) \sim n^{-1/\gamma d^*} \tag{2.62}$$

where d^* is the design complexity and equals to 1 when all components are independent.

Search models successfully explain the emergence of power-law scaling in repeated attempts and serve as the basis of our frameworks (e.g. $k \to \infty$ limit). Yet they cannot account for the co-existence of two groups and their diverging patterns.

2.7.5. Individual learning models

There has also been an active line of inquiry in explaining practice curves in individual tasks [39, 40, 97, 98]. These models use psychology models as well as cognitive theories to explain 'practice makes perfect'.

2.7.5.1. Newell and Rosenbloom's chunking model. To explain the power-law scaling observed in human task performance, e.g. inverted text reading and ten-finger game, Newell and Rosenbloom modeled the learning process using chunking theory [39]. In this model, there is a tree structure for goal hierarchies of height H and the speed-up of task completing is due to the emergence of higher-order chunks. The current highest order of chunk is denoted as η, leading to

$$\frac{dT}{dN} = \frac{dT}{d\eta}\frac{d\eta}{dN} \tag{2.63}$$

The model further assumes each non-terminal goal has β non-terminal subgoals and ω terminal subgoals. As one constructs chunks of higher levels, the corresponding time to perform a new attempt decreases exponentially following

$$\frac{dT}{d\eta} \sim \beta^{H-\eta} \tag{2.64}$$

If we also assume the chunking rate is linear with respect to time and the birth of a single level-h chunk requires time $s(h)$, we have

$$\frac{d\eta}{dN} \sim \frac{\beta^{\eta-H}}{s(\eta)} T \tag{2.65}$$

Therefore, if $s(\eta)$, the number of possible states for goals at level η (complexity at this level), takes an exponential form as $s(\eta) \sim e^{\alpha\eta}$, which is consistent with the tree structure, we have

$$\frac{dT}{dN} \sim \frac{(T+E)^{-x}}{T} \tag{2.66}$$

which follows a power law scaling. Hence, by combining two exponential forms in a tree structure, the model successfully derives the power law scaling.

2.7.5.2. Anderson's model. Based on ACT's strengthening process, Anderson developed a model explaining cost decay [40]. The model assumes the amount of practice as S and the production execution in ACT takes the form

$$T = c + aS^{-1} \tag{2.67}$$

The amount of past practice also decays as a power law of practice time:

$$S = \sum_{i=0}^{P-1} s(i,P) \sim \sum_{i=1}^{P} i^{-d} \sim P^{1-d} \tag{2.68}$$

Therefore, we have

$$T = C' + A'P^{d-1} \tag{2.69}$$

The two models are very relevant to our settings and can predict power law temporal scaling in the successful group. They represent two fundamental classes of cognitive architectures in related studies: ACT and Soar (and their variants) [99], highlighting the role of memory and chunks in learning process. Yet such mechanisms are more appropriate for modeling simple tasks rather than complex innovative ones and cannot account for the co-emergence of success and unsuccessful groups.

2.7.6. Urn models

Urn model and its variants are among the canonical models in social physics as well as innovation process [100]. This model family is closely related to the famous Heaps' law [50], originally predicting that the number of distinct words S in a paragraph of length n scales as

$$S(n) \sim n^{\beta},\ 0 < \beta \leq 1 \tag{2.70}$$

Note that if we assume generating a new word costs unit time, we know the expected time spent on the n-th 'word' follows

$$t_n \sim n^{\beta-1} \equiv n^{-\gamma},\ 0 < \gamma \leq 1 \tag{2.71}$$

which recovers our empirical findings. Here we review several generative models explaining this scaling.

2.7.6.1. Simon's model. Simon's model is among the earliest frameworks modeling 'cumulative advantages' [93]. It assumes that (1) There is always constant probability p for an agent to take a new word for the next element; (2) Otherwise (with probability $1-p$) the agent reuses past words based on frequency, i.e. randomly select a word from the past sequence. This model predicts a linear scaling between S and n i.e. $\beta = 1$, which can only explain the emergence of the unsuccessful group.

2.7.6.2. Tria's model. By extending studies on urn model, Tria et al [94] assume an urn U of ideas and a sequence of S to generate. Every time an element is sampled from U to S, ρ copies are put back to U. Further, if this sampled idea is new in S, it triggers ν adjacent new ideas, hence the number of different ideas in a sequence follows the master equation

$$\frac{dD}{dt} \approx \frac{\nu D}{\rho t + (\nu + 1)D} \tag{2.72}$$

The solution reveals that D grows linearly with t for $\nu > \rho$, but follows Heaps' law $D \sim t^{\nu/\rho}$ for $\nu < \rho$. These predictions are similar to the first phase transition point k^* in our model.

2.7.6.3. Iacopini's model. To further document the impact of past transition sequence in innovative attempts, a recent paper [95] proposed a network-based model, where ideas are represented as nodes, and one can travel from one idea to another when they are linked by a weight. The process is set to be a weighted random walk on networks, following

$$P^t(i \to j) = \frac{w_{ij}^t}{\sum_k w_{ik}^t} \tag{2.73}$$

When a specific path $i \to j$ is traveled, the weight of this edge is updated

$$w_{ij}^{t+1} = w_{ij}^t + \delta w \tag{2.74}$$

Depending on different network structures, the model can lead to scaling $S \sim n^\beta$ with varying β.

While this class of models does not capture the performance dynamics underlying failures, they are highly relevant to our study in that their predictions are consistent with the temporal patterns observed in our data.

2.7.7. Other models

2.7.7.1. Levy's model. Levy modeled the improvement of productivity based on the limited range of output denoted as P [82]. Given the current rate of production after producing q items, $Q(q)$, the improvement of production rate is proportional to the amount that the process can improve, i.e.

$$\frac{dQ(q)}{dq} = \mu[P - Q(q)] \tag{2.75}$$

leading to

$$Q(q) = P[1 - e^{a+\mu q}] \tag{2.76}$$

Levy's model captures a kind of production process where the final plateau part is significant, but it fails to predict the power-law form of productivity improvement.

2.7.7.2. Shrager's model. By collecting and analyzing data of path length in the bit game, Shrager et al developed a graph-dynamic model for route-finding in ER networks $G(n,p)$ [87]. The authors proposed a strategy where the individual randomly selects an edge after deleting the ones moving away from the destination with probability r. The number of trials increases the network density p linearly and the cost is the path length of the whole process s. For r near 0, the model predicts

$$s \sim \frac{2}{p}(1-r)^{\ln n/\ln(np)} \tag{2.77}$$

while for r near 1, the model predicts

$$s \sim \ln n/\ln(np) \tag{2.78}$$

2.7.7.3. Sahal's model. Sahal explains the progress function in industry productions through probabilistic and deterministic models [85]. The model assumes different manpower levels and $X(s,t)$ to be the number of product quantities requiring s amount of manpower at time t. If we assume the improvement across u manpower levels does not depend on the current level and can be formulated as $p(u)$, yielding

$$X(s,t+1) = \sum_{u=-n}^{1} X(s-u,t)p(u) \tag{2.79}$$

If we define $X(s) = \lim_{t\to\infty} X(s,t)$, the solution of this equation can be formulated as

$$X(s) = b^s,\ 0 < b < 1 \tag{2.80}$$

The model further assumes levels manpower are distributed on a logarithmic scale with width h, obtaining

$$F(Y) \sim Y^{-\log b/h} \tag{2.81}$$

where $F(Y)$ is the number of product quantities requiring manpower greater than Y.

2.7.7.4. Johnson's model. Johnson et al reported a similar scaling from the time interval of terrorist attacks and other human confrontations [37]. An illustrative model for this scenario considers confrontation between 'Red Queen' and 'Blue King', and the advantage of Red Queen after n events, $R(n)$, can be formulated as

$$R(n) = \sum_{i=1}^{n} x_i \tag{2.82}$$

where x_i takes value $+d$ or $-d$ with probability 1/2. Depending on the auto-correlation of x_i, one can get

$$R(n) \sim n^b d, 0 \leq b \leq 1 \tag{2.83}$$

Taking the inverse of the advantage, we get the attack rate scales as a negative power law of n, i.e.

$$\tau_n \sim n^{-b}, 0 \leq b \leq 1 \tag{2.84}$$

2.7.7.5. Clauset's model. Clauset's model [88] also predicts the temporal pattern of terrorist attacks, but in a very different way from Johnson's model [37]. Indeed, if we assume that the size of terrorism organizations scales linearly with its past attacks, i.e.

$$s(n+1) = s(n) + \eta \tag{2.85}$$

The model further assumes a new takes time as the inverse of organization size, i.e.

$$\Delta t \sim 1/s \tag{2.86}$$

Taken together, we have

$$\Delta t \sim 1/n \tag{2.87}$$

This model successfully links group size to temporal dynamics, predicting a power law scaling. Yet it only applies to group dynamics and the exponent of power law in the original linear assumption is restricted to be -1.

One commonality among these models is that they lack predictions of the interplay between performance and time. By contrast, our data show that the temporal scaling cannot be simply explained by agents optimizing time cost t_n since the performance also improves for the successful group. These models also cannot explain the co-existence of success and unsuccessful groups observed in our data.

2.7.8. Summary of contributions

As we will show in this chapter, despite the ubiquity of power laws across a wide variety of settings [20–22, 51, 101, 102] and the foundational literature on learning curves [26, 32, 39–42, 84, 97], none of the existing models, to our knowledge, anticipated the existence of the early signals documented in the paper (Table 2.3). As such, the paper makes several contributions which we next summarize in terms of its empirical measurements, theoretical contributions, and predictive signals:

(1) Empirical contributions: Our quantitative understanding of the dynamics of failure is important, but has remained limited, due to difficulties in collecting large-scale datasets that capture failures. This highlights the first contribution of our paper – to be able to assemble large datasets from three disparate domains that contain records of both success and failure cases.

(2) Theoretical contributions: These new datasets allow us to derive among the first empirical evidence about the dynamics of failure to test existing models. In particular, the simplicity of measurements in Figure 2.1 highlights the fundamental tension with existing modeling framework and the paucity of quantitative approaches thus far to model failures. In this paper, By establishing a new theoretical basis for understanding failures, our paper not only explains empirical patterns that existing models cannot capture, but also predict new patterns that existing models did not anticipate (Fig. 2.4). As such, the model is unique in its ability to (i) predict two fundamentally different behaviors simultaneously at two extremes (e.g., $k = 0$ and $k = \infty$), hence serving as the first model to unify existing paradigms; and (ii) reveal a highly discontinuous pattern between progression and stagnation regimes. This further leads to four new predictions, all of which are tested and validated across our three datasets. This was only possible thanks to the new theoretical insights, and in particular the novel predictions that our model offers.

(3) Predictive signals: Our findings unveil identifiable yet previously unknown early signals that allow us to distinguish failure dynamics that will lead to ultimate victory or defeat. Traditionally the primary distinction between ultimate victory and defeat has been attributed to differences in luck, learning strategies or individual characteristics, but here our model offers an important new explanation with crucial implications: Even in the absence of distinguishing initial characteristics, agents may experience fundamentally different outcomes. As such, our model shows that the success and unsuccessful groups may be initially similar, but each follows their respective, highly predictable patterns, distinguishable long before the eventual outcome becomes apparent. Specifically, we show that observing the timing of each attempt alone can help us identify those more likely to succeed. Considering the myriad factors related to success in a grant proposal/startup company/terrorist attack, this level of predictive power achieved by a singular, simple predictor is somewhat unexpected.

2.8. Concluding remarks

Together, these results support the hypothesis that if future attempts systematically build on past failures, the dynamics of repeated failures may reveal statistical signatures discernible at an early stage. Traditionally the main distinction between ultimate success and failure following repeated attempts has been attributed to differences in luck, learning strategies or individual characteristics, but here our model offers an important new explanation with crucial implications: Even in the *absence* of distinguishing initial characteristics, agents may still experience fundamentally different outcomes. These results not only deepen our understanding of the complex dynamics beneath failure, they also hold lessons for individuals and organizations that experience failure and the institutions that aim to facilitate or hinder their eventual breakthrough.

CHAPTER 3

Coevolution between policy and science during the pandemic

Disconnects between science and policy, where important scientific insights may be missed by policymakers and bad scientific advice may infect decision making, are a long-standing concern [103–109]. Yet our systematic understanding of the use of science in policy remains limited [103, 106–108], partly due to the difficulty in reliably tracing the co-evolution of policy and science at a large, global scale [105]. Today, the world faces a common emergency in the COVID-19 pandemic, which presents a dynamic, uncertain, yet extraordinarily consequential policy environment across the globe. Here we combine two large-scale databases capturing policy and science and their interactions, allowing us to examine the co-evolution of policy and science during the pandemic. Our analysis suggests that many policy documents in the COVID-19 pandemic substantially access recent, peer-reviewed, and high-impact science. And policy documents that cite science are especially highly cited within the policy domain. At the same time, there is notable heterogeneity in the use of science across policy-making institutions. The tendency for policy documents to cite science appears mostly concentrated within intergovernmental organizations (IGOs), such as the World Health Organization (WHO), and much less so in national governments, which consume science largely indirectly through the IGOs. This close co-evolution between policy and science offers a useful indication that a key link is operating, but it has not been a sufficient condition for effectiveness in containing the pandemic.

The rapid production of new science during COVID-19 raises key questions about its use in policy during the pandemic. There is long-standing skepticism over connections between science and policy, which are often thought to be highly disconnected spheres. For example, the “two communities” theory in knowledge utilization [109] highlights a substantial gap between scientists and policy makers, disconnecting research from the policy process. Related viewpoints suggest that policy makers may not be able to distinguish relatively robust scientific ideas from less established ones [104]. Particularly in the pandemic setting, there is substantial concern that policy may take up non-vetted and potentially incorrect scientific results. For example, preprint servers have played an outsize role in disseminating COVID-19 related research [110]. While open science greatly facilitates the sharing of data and research [110] and allows the wider community to check and interrogate the results and claims, publicly releasing science before it passes peer review may undermine the rigor of scientific evidence accessible to the public [111]. In the age of misinformation, this may create enduring harms if the evidence presented turns out to be less robust. Such concerns are further heightened by examples of widely-reported and then retracted results regarding COVID-19 [112].

To explore COVID-19 science and policy, we harness a novel, large-scale database, Overton, which records policy documents sourced globally from government agencies, think tanks, and IGOs. For each policy document, we then match scientific references to our second dataset, Dimensions, a large-scale publication and citation database, offering a unique opportunity to examine the role of science in the global policy response to COVID-19.

3.1. Data Description

3.1.1. Overton policy data

To understand the global policy response to the COVID-19 pandemic, we leverage a novel dataset provided by Overton (https://www.overton.io). Overton data captures among the world's largest collection of policy documents [113]. Policy documents are broadly defined as documents written primarily for or by policymakers, and include documents from government agencies, think tanks as well as intergovernmental organizations (IGOs) [113]. The data is updated weekly, allowing us to trace how policy responses evolve in nearly real time. By the time of our data collection, the Overton database had captured over 43,000 policy documents from 114 countries in 2020, collecting documents from more than 1,200 different sources worldwide.

In this paper, we use the full set of policy documents published from January 1st 2020 to May 26th 2020. We use an API to obtain policy documents from each policy source separately. For each document, we have information on its title, original URL, publication date, document type, policy source, and subject classification codes.

To identify COVID-19 related policy documents, we leverage Overton's technical capabilities which combine translation of policy documents into English with keyword-based search for COVID-19 related keywords across multiple languages [114]. While the data cover a large number of countries, there are notable exceptions. First is mainland China. Although China experienced the COVID-19 outbreak early on in 2020, policy documents from mainland China are missing in the Overton database for 2020. Personal correspondence with the Overton team has not pinpointed the reason for this missing data, but it may have to do with web crawling issues on Chinese government websites [115]. Second is the Netherlands. Our access to the API only allows for at most 2,000 results per query, and the Netherlands exceeded this limit with over 4,000 documents. We therefore exclude Netherlands from our analysis to avoid potential bias.

We focus on policy documents published by (i) governmental agencies and think tanks in country members of the United Nations, and (ii) intergovernmental organizations (defined by “IGO” and “EU” source labels in the Overton data). For the temporal coverage, we froze our data on May 26th, 2020. We noticed that there is an unusual number of documents published on January 1st 2020 (1,036 documents, about 4 times of a normal day), possibly due to default classification to this date. We further excluded these documents from our analyses to avoid including policy documents that have inaccurate publication dates. Lastly, since our main focus is on policy documents, we follow Overton’s suggestion [115] and further filter on the document type by using only “publications” (95% of the total documents), removing other types such as “working papers”, “transcripts”, “blog posts”, and “clinical guides”.[1]

In total, we analyzed 7,730 COVID-19 documents out of 37,725 documents published across 114 countries (including 402 think tanks) and 55 IGOs (including 4 in EU). The world map of the policy data coverage is shown in Fig. 3.1.

[1] Note that some World Health Organization (WHO) documents (under the “publications” category) can be characterized as “interim guidance” documents, but it seems they are not the same kind of clinical guidelines as commonly defined in PubMed. Indeed, the WHO website describes their guidance on COVID-19 as “*meant for health decision makers who adapt the information for their country and context*”. Upon closer inspection, we find that these documents differ significantly from the formal clinical guidelines in repositories like PubMed, and are characterized by shorter length, more general recommendations, and fewer scientific figures/tables. While these documents may not be aimed at all public policymakers, they tend to be directly consumed by policymakers such as the CDC. For these reasons, in this paper we consider these documents as policy documents, as they seem relevant not only for health workers but also policymakers in public health.

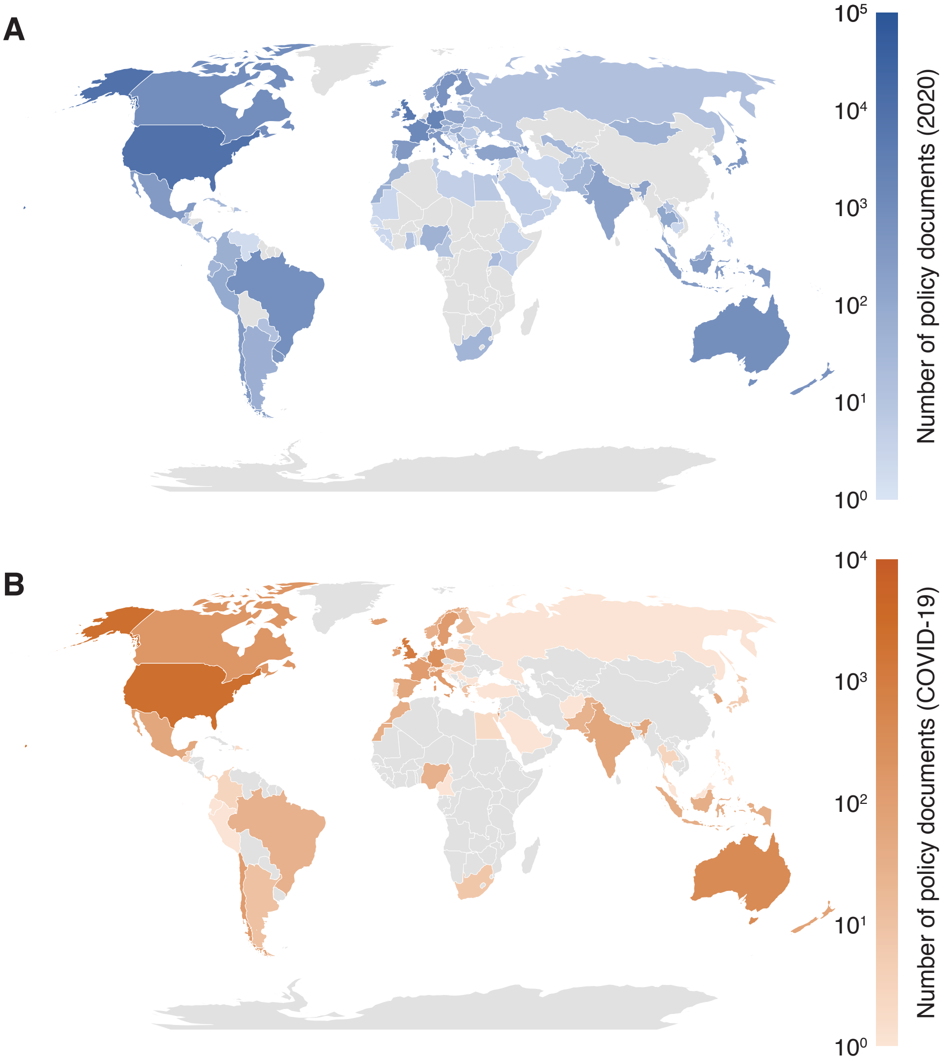
A
10^5
10^4
10^3
10^2
10^1
10^0
Number of policy documents (2020)
B
10^4
10^3
10^2
10^1
10^0
Number of policy documents (COVID-19)

Figure 3.1. **World map of policy data coverage.**

(A) Number of policy documents published in 2020 in our data. (B) Number of COVID-19 policy documents. We visualize (#docs+1) for the logarithmic color scale.

3.1.2. Dimensions publication data

We further link the scientific references cited within the policy documents to a database of scientific publications. For this database, we use Dimensions [116], a data product by

Digital Science. Dimensions is one of the largest citation databases, including over 100 million publications accessible from journals, conference proceedings, books and chapters, and preprint servers. Publication data are updated on a daily basis and accessible from API, allowing us to collect reference information in a timely manner. Each scientific reference from Overton has a unique DOI (Digital Object Identifier), one of the most commonly used identifiers for scientific publications. We retrieve papers from the Dimensions API using the DOI information. We find that the vast majority of the references can be matched to Dimensions records (79,669 out of 84,964 papers, or 93.8% of the scientific references). For each paper we obtain information on its title, author list, affiliation(s), publishing venue, publication date, fields of study, references and citations received. Among the papers we analyzed, 79 of the 79,669 are matched to more than one record in Dimensions; hence to avoid duplications we keep the single item with the most complete records, as determined by the number of references and citations.

We also constructed another set of COVID-19 related scientific publications by searching for papers published in 2020 with the following query suggested by Dimensions:

“2019-nCoV” OR “COVID-19” OR “SARS-CoV-2” OR “HCoV-2019” OR “hcov” OR “NCOVID-19” OR “severe acute respiratory syndrome coronavirus 2” OR “severe acute respiratory syndrome corona virus 2” OR ((“coronavirus” OR “corona virus”) AND (Wuhan OR China OR novel)),

yielding in total 40,732 papers published in 2020 out of all articles indexed by Dimensions with unique DOIs, among which 2.3% (933 papers) have been cited by COVID-19 policy documents.

3.1.3. COVID-19 case and death tracking data

We use country-level daily statistics for COVID-19 confirmed cases and deaths from the COVID-19 Data Repository that is maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [117]. The data has been extensively used in COVID-19 related studies. We froze the data of cases and deaths on May 30th 2020, and the data used in our analysis covers a period from January 22nd to May 26th 2020.

3.1.4. Comparing data coverage between Overton and Altmetric

Tracking how science is cited in policy is important for a large variety of issues in science and science policy, but it remains a long-standing challenge [105, 118–122]. There have been several initial attempts at collecting this data, most notably by Altmetric, a company under Digital Science's portfolio. Prior studies on policy-science citations have examined the overall coverage and reliability of the Altmetric data. For example, Haunschild and Bornmann [105] offered an early systematic study of Altmetric data, recommending further improvements on policy-related sites tracked by Altmetric and cautions on the use of this data for a systematic analysis of policy-science linkages. Later, Tattersall and Carroll [119] also pointed out several data quality issues, including false attribution of authors and affiliations as well as unclear identification of policy documents. Recently, Yu et al. [120] performed a manual validation on 2,079 records from the Altmetric data, arguing that errors on policy citation to science may be "relatively minor" but reasons for such errors "remain to be further investigated". Researchers have also used Altmetric data to quantify the mentions of scientific publications by policy across different fields

[105, 123] and research institutions [119, 124]. Recent studies have also highlighted the fact that policy to science citations could serve as an important indicator for societal impact [123, 125]. For a more detailed overview of the Altmetrics literature, readers may refer to the recent review paper by Tahamtan and Bornmann [126].

Motivated by the Altmetrics literature, we next systematically compare the data coverage of Overton with Altmetric. Here we further supplement the Altmetric data with Dimensions, a sister company under the same portfolio company Digital Science. We combine both Altmetric and Dimensions as the expanded Altmetric data, and compare it with our Overton data. First, we find that there are fundamental differences in data coverage between Overton and the expanded Altmetric data:

(i) Overton sources from an order of magnitude more policy sources: Altmetric tracks policy documents from 131 sources, while Overton tracks over 1,200 various policy sources, including many unique yet important governmental institutions (e.g., the US White House, the House Committees, Australian Government Department of Health, India National Centre for Disease Control, and many national/state governments) and think tanks (e.g., the RAND corporation, and the Heritage Foundation).

(ii) Overton tracks multiple times more policy documents and their linkages with science: Overton covers 6.5 times more policy documents than Altmetric/Dimensions (e.g., 2.94M in Overton vs. 452K in Altmetric by the end of 2019). As a result, Overton covers more than twice as many policy-science linkages than Altmetric (e.g., the Altmetric dataset includes 1.28M scientific papers cited by policy documents, whereas Overton currently includes 2.84M).

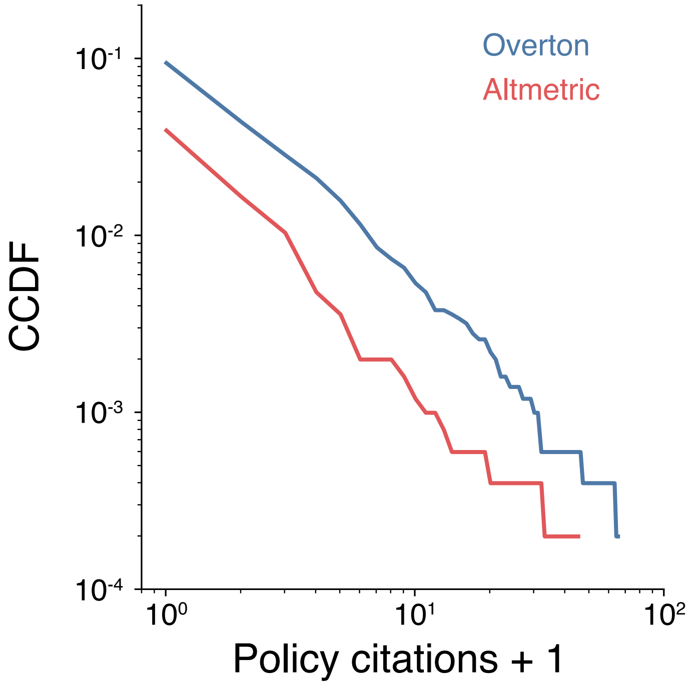

Figure 3.2. **Policy citation counts in Overton and Altmetric for 5,000 COVID-19 papers.**

Overton covers an order of magnitude more citation linkages than Altmetric.

(iii) Beyond the differences in sheer counts, there's a categorical difference in data coverage between Overton and other existing datasets. For example, compared with Altmetric, Dimensions, and PlumX, Overton is, to the best of our knowledge, the only set that includes policy-to-policy citation information, which enables many of our analyses in the main text.

Next, we show Overton data appears to represent a qualitative leap from the state-of-the-art datasets in quantifying science-policy linkages.

(1) There is a fundamental difference in counting policy-science linkages. To provide a direct comparison, here we randomly select 5,000 COVID-19 papers and track their

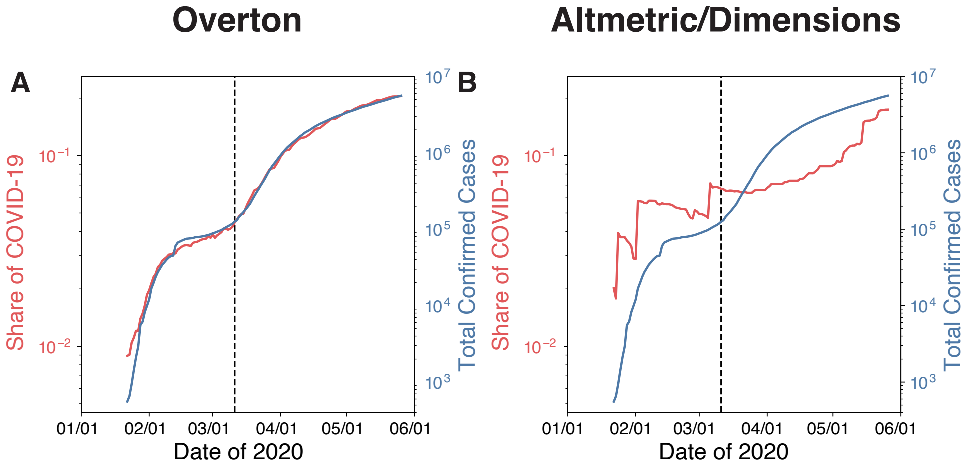

Figure 3.3. **Relationship between the share of COVID-19 policy documents among all policy documents and the number of total confirmed cases.**

(A) Overton data shows a much higher goodness of fit (0.705) with total confirmed cases than (B) Altmetric data (0.231) does, as measured by the cross-correlation of the two time-series and as can be seen visually. Throughout the figures, the black vertical dashed line marks the date of the WHO's pandemic declaration.

policy citations in the Overton and Altmetric databases using their API access. The two distributions shown in Fig. 3.2 suggest that Overton covers an order of magnitude more citation linkages than Altmetric.

(2) There is a qualitative difference in how well the data corresponds to facts on the ground. To show this, we use our expanded Altmetric/Dimensions data to repeat our main findings – to understand to what degree the evolution of COVID-19 policy documents corresponds to the reality of the pandemic. We use the date information contained in the Dimensions database to construct the temporal evolution of policy documents. While in the Overton data, the number of COVID policy documents closely tracks the course of the pandemic (Fig. 3.3A), the numbers recorded in Altmetric/Dimensions data are only weakly related (Fig. 3.3B), if at all to the case counts.

Together these results indicate that Overton represents the state-of-the-art dataset capturing policy documents and scientific references at a new level of scale and comprehensiveness. The remarkable consistency between COVID policy attention and case counts documented in this paper offers additional assurance, showing Overton data's unique ability to capture up-to-date information in a way that is consistent with facts on the ground.

3.1.5. Independent validations of the Overton data

As should become clear in our analysis above, the difficulty in validating the Overton data lies in the fundamental limitations of other existing datasets, which are not adequate to assess the coverage and reliability of our data. Nonetheless, here we proceed with two types of analyses. First, we bring in a novel, large-scale dataset to independently validate the Overton data, and show close consistencies across the two datasets. Second, we repeat the analyses that can be performed on the Altmetric data to cross-validate our findings, and show that they uncover consistent patterns. Here we describe our first analysis, and we present the second analysis together with other robustness checks in the following sections.

We bring in a novel, large-scale dataset that contains an independent collection of policy documents and their citations to scientific publications, which were all collected using completely different methodologies. More specifically, we partnered with Microsoft to leverage the Microsoft Bing search engine to collect over 6 million government documents available online across all branches of the U.S. government. We developed a machine reading technology to systematically identify academic publications that are referenced in these government documents, using the technologies at the Microsoft Academic Graph (MAG), and matching these references to the MAG, which is one of the largest bibliometric databases in the world. This novel pipeline allows us to collect a large dataset on how government documents consume scientific knowledge within the United States.

To systematically compare the science-policy linkages, we performed two different measurements for papers published across all scientific fields recorded in the MAG: (i) We first measure the relative chance, by MAG field, for papers to be cited by US government documents, and compare each field across the two datasets. (ii) We measure the probability for a policy-cited paper to be a hit paper within science, defined as being in the top 1% of scientific citations within the same field and year, and compare this quantity between the two datasets, to understand if they cover a similar proportion of high-quality science. We find that, although the two datasets are collected for different purposes using different approaches and technologies, the measurements carried out independently across the two datasets show remarkable consistencies (Fig. 3.4). These results further cross-validated the reliability of the Overton data.

3.1.6. Data examples and discussions

To illustrate how science is used by the policy documents recorded in our data, here we provide a few examples of COVID-19 policy documents in Overton and their citations to science as case studies. For illustration purposes, we sampled policy documents with both high and low policy citations. We find that while policy may use science for a variety of purposes, a majority of these uses recorded in our data seem to show substantial

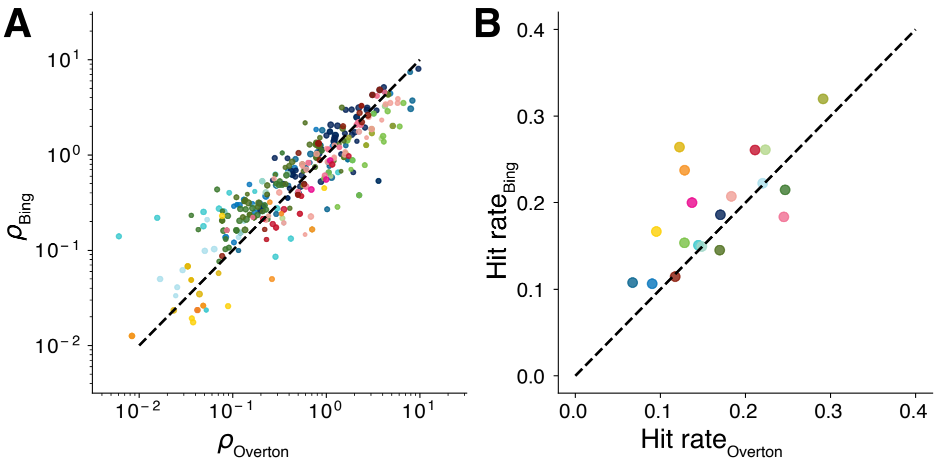

Figure 3.4. **Comparing Overton data with Bing data.**

(A) For 294 level-1 fields, we measure the relative chance for papers in a MAG field to be cited by US government documents. (B) For 19 level-0 fields, we measure the probability for a policy-cited paper to be a hit paper, defined as a paper in the top 1% of scientific citations within the same field and year. For both quantities, measurements carried out independently across the two datasets show substantial consistency.

consistency between the topics of policy and their scientific references. These examples are illustrated below:

Policy document Example 1: *Infection prevention and control during health care when novel coronavirus (nCoV) infection is suspected.*

- Source: World Health Organization [Type: IGO, 76 citations from COVID-19 policy documents]
- Excerpt: [*Some aerosol generating procedures have been associated with increased risk of transmission of coronaviruses (SARS-CoV and MERS-CoV) such as tracheal intubation, non-invasive ventilation, tracheotomy, cardiopulmonary resuscitation, manual ventilation before intubation and bronchoscopy 7. . . . Environmental and engineering controls . . . . Both controls can help reduce the spread of many pathogens during health care 10.*]
- Science referenced:

[7] Hui, D. S. (2017). Epidemic and emerging coronaviruses (severe acute respiratory syndrome and Middle East respiratory syndrome). Clinics in Chest Medicine, 38, 71-86.

[10] Jefferson T, Del Mar CB, Dooley L et al. Physical interventions to interrupt or reduce the spread of respiratory viruses. Cochrane Database of Systematic Reviews, 2011, 7:CD006207.

Policy document Example 2: *Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief, 27 March 2020.*

- Source: World Health Organization [Type: IGO, 8 citations from COVID-19 policy documents]
- Excerpt: [*There are reports from settings where symptomatic COVID-19 patients have been admitted and in which no COVID-19 RNA was detected in air samples.10-11 In addition, it is important to note that the detection of RNA in environmental samples based on PCR-based assays is not indicative of viable virus that could be transmissible.*]
- Science referenced:

[10] Cheng, V. C., Wong, S. C., Chen, J. H., Yip, C. C., Chuang, V. W., Tsang, O. T., ... & Yuen, K. Y. (2020). Escalating infection control response to the rapidly evolving epidemiology of the Coronavirus disease 2019 (COVID-19) due to SARS-CoV-2 in Hong Kong. Infection Control & Hospital Epidemiology, 41(5), 493-498.

[11] Ong, S. W. X., Tan, Y. K., Chia, P. Y., Lee, T. H., Ng, O. T., Wong, M. S. Y., & Marimuthu, K. (2020). Air, surface environmental, and personal protective equipment contamination by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from a symptomatic patient. JAMA, 323(16), 1610-1612.

Policy document Example 3: *Rapid Risk Assessment: Coronavirus disease 2019 (COVID-19) in the EU/EEA and the UK– ninth update.*

- Source: European Centre for Disease Prevention and Control [Type: IGO, 2 citations from COVID-19 policy documents]
- Excerpt: [*In an analysis of data from a cohort of patients with COVID-19 and a metaanalysis of findings from publications, viral RNA was detected in stool samples from 48.1% (95% CI, 38.3%–57.9%) of the patients—even in stool collected after the respiratory samples tested negative [37]. It should be noted that detection of viral RNA by PCR does not equate with infectivity.*]
- Science referenced:

[37] Cheung, K. S., Hung, I. F., Chan, P. P., Lung, K. C., Tso, E., Liu, R., ... & Yip, C. C. (2020). Gastrointestinal manifestations of SARS-CoV-2 infection and virus load in fecal samples from the Hong Kong cohort and systematic review and meta-analysis. Gastroenterology. 59(1):81-95.

Policy document Example 4: *COVID-19 in racial and ethnic minority groups.*

- Source: Centers for Disease Control and Prevention, USA [Type: government, 1 citation from COVID-19 policy documents]
- Excerpt: [*Research also suggests that racial residential segregation is a fundamental cause of health disparities. For example, racial residential segregation residential segregation is linked with a variety of adverse health outcomes and underlying health conditions 2-5*].
- Science referenced:

[2] Bravo MA, Anthopolos R, Kimbro RT, Miranda ML. Residential racial isolation and spatial patterning of type 2 diabetes mellitus in Durham, North Carolina. Am J Epidemiol 2018;187(7):1467–7.

[3] Anthopolos R, James SA, Gelfand AE, Miranda ML. A spatial measure of neighborhood level racial isolation applied to low birthweight, preterm birth, and birthweight in North Carolina. Spat Spatio-Temporal Epidemiol 2011;2(4):235–46.

[4] Hearst MO, Oakes JM, Johnson PJ. The effect of racial residential segregation on black infant mortality. Am J Epidemiol 2008;168(11):1247–54.

[5] Jackson SA, Anderson RT, Johnson NJ, Sorlie PD. The relation of residential segregation to all-cause mortality: a study in black and white. Am J Public Health 2000;90(4):615–7.

Policy document Example 5: *Symptom-Based Strategy to Discontinue Isolation for Persons with COVID-19: decision memo.*

- Source: Centers for Disease Control and Prevention, USA [Type: government, 0 citations from COVID-19 policy documents]
- Excerpt: [*Viral burden measured in upper respiratory specimens declines after onset of illness (CDC unpublished data, Midgely 2020, Young 2020, Zou 2020, Wölfel 2020).*]
- Science referenced:

Midgley CM, Kujawski SA, Wong KK, Collins, JP, Epstein., Killerby ME et al. (2020). Clinical and Virologic Characteristics of the First 12 Patients with Coronavirus Disease 2019 (COVID-19) in the United States. Nature Medicine, in print.

Young BE, Ong SWX, Kalimuddin S, Low JG, Ta, SY, Loh J, et al. (2020). Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore. JAMA.

Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. (2020). SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. N Engl J Med, 382(12), 1177-1179.

Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature.

Policy document Example 6: *PHLN statement on use of saliva as an alternative specimen for the diagnosis of SARS-COV-2.*

- Source: Australian Government Department of Health, Australia [Type: government, 0 citations from COVID-19 policy documents]
- Excerpt: [*PHLN continues to monitor the emerging literature with regard to the performance of saliva collection as alternative specimen for use in PCR testing. Some early validation studies have been conducted both internationally and in Australia, which indicate promising results.1-2*]. Notes: Public Health Laboratory Network (PHLN).
- Science referenced:

[1] Khurshid, Z., Zohaib, S., Joshi, C., Moin, S. F., Zafar, M. S., & Speicher, D. J. (2020). Saliva as a non-invasive sample for the detection of SARS-CoV-2: a systematic review. medRxiv.

[2] Azzi, L., Carcano, G., Gianfagna, F., Grossi, P., Dalla Gasperina, D., Genoni, A., ... & Maurino, V. (2020). Saliva is a reliable tool to detect SARS-CoV-2. Journal of Infection. 81(1), e45-e50.

Policy document Example 7: *Guidelines for disinfection of quarantine facility (for COVID-19).*

- Source: National Centre for Disease Control, India [Type: government, 0 citations from COVID-19 policy documents]
- Excerpt: [*According to studies assessing the environmental stability of other coronaviruses, the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) is estimated to survive several days in the environment and the Middle East Respiratory Syndrome-related coronavirus (MERS-CoV) more than 48 hours at an average room temperature (20°C) on different surfaces [1-3]*]
- Science referenced:

[1] Van Doremalen, N., Bushmaker, T., & Munster, V. J. (2013). Stability of Middle East respiratory syndrome coronavirus (MERS-CoV) under different environmental conditions. Eurosurveillance, 18(38), 20590.

[2] Otter, J. A., Donskey, C., Yezli, S., Douthwaite, S., Goldenberg, S., & Weber, D. J. (2016). Transmission of SARS and MERS coronaviruses and influenza virus in healthcare settings: the possible role of dry surface contamination. Journal of Hospital Infection, 92(3), 235-250.

Policy document Example 8: *Lineamiento para la prevención y mitigación de COVID-19 en la atención del embarazo* [Guidelines for the prevention and mitigation of COVID-19 in the care of pregnancy]

- Source: Government of Mexico, Mexico [Type: government, 0 citations from COVID-19 policy documents]
- Excerpt: [*No se ha confirmado la transmisión vertical toda vez que las muestras de líquido amniótico, tejido placentario, sangre de cordón umbilical y exudado faríngeo en los recién nacidos fueron negativas en las series de casos publicadas hasta ahora (Chen et al, 2020; Zhu et al, 2020;Schwartz, 2020)*]

Machine translation to English [*Vertical transmission has not been confirmed since the samples of amniotic fluid, placental tissue, umbilical cord blood and pharyngeal exudate in newborns were negative in the case series published so far (Chen et al, 2020;Zhu et al, 2020;Schwartz, 2020)*]

- Science referenced:

Chen, H. et al (2020) Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. Lancet 2020;395: 809–15.

Zhu, H. et al (2020) Clinical analysis of 10 neonates born to mothers with 2019-nCoV Pneumonia. Transl Pediatr 2020;9(1):51-60

Schwartz, D. (2020) An Analysis of 38 Pregnant Women with COVID-19, Their Newborn Infants, and Maternal-Fetal Transmission of SARS-CoV-2: Maternal Coronavirus Infections and Pregnancy Outcomes. Archives of Pathology & Laboratory Medicine.

Policy document Example 9: *Public health principles for a phased reopening during COVID-19: Guidance for governors - AEI.*

- Source: American Enterprise Institute [Type: think tank, 0 citations from COVID-19 policy documents]
- Excerpt: [*Other studies that attempt to reconstruct transmission chains among confirmed cases have also found that prolonged close contact is the source of most new infections. Some special settings have also been identified.*]
- Science referenced:

Pung, R., Chiew, C. J., Young, B. E., Chin, S., Chen, M. I., Clapham, H. E., ... & Low, M. (2020). Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. Lancet. 395(10229), 1039-1046.

Note that policy-science citations may occur for different reasons [108, 127], including (i) instrumental uses (knowledge directly applied to solve problems); (ii) conceptual uses (research influences or informs the way policymakers think); (iii) tactical uses (citing research to support or challenge an idea) among others, suggesting the need to understand the semantics of the policy-science citations. While our dataset offers among the largest collection of policy documents and their linkages with science, it is not yet possible to evaluate the semantics behind these citations at scale, as also pointed out in the NAS report [103] and other studies [128]. Nevertheless, we find many of the science citations in Overton policy documents can be quite relevant. For example, we discover a high degree of consistency between the topics of the policy documents and the science they cited,

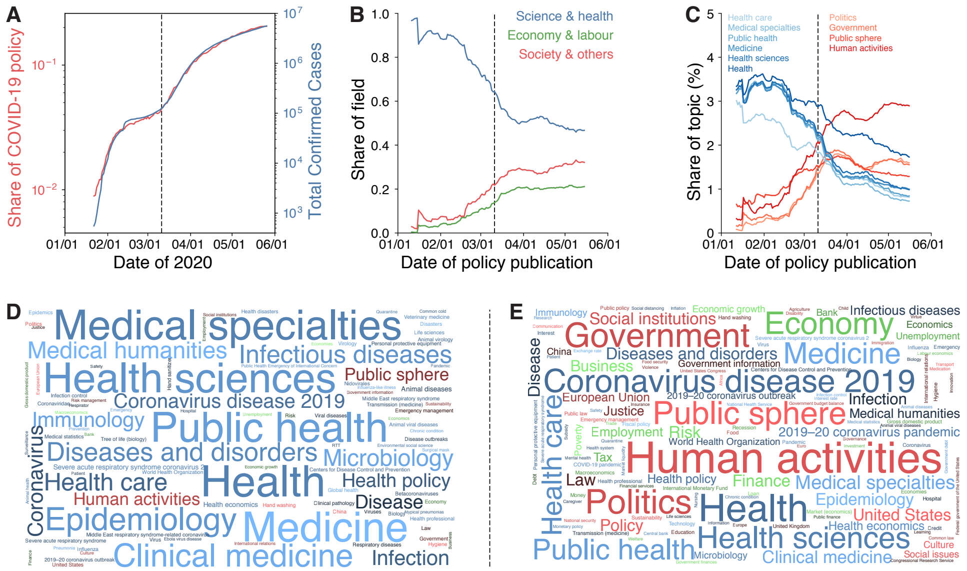

Figure 3.5. **The evolution of policy during the COVID-19 pandemic.**

(A) Policy documents mirror the case dynamics, showing a synchrony between the share of COVID-19 policy documents among all policy documents and the number of total confirmed cases. (B) The share of COVID-19 policy documents across three broad subject categories (21-day moving average). (C) The share of COVID-19 policy documents across topics (21-day moving average). Color blue and red marks health- and economy-related topics, respectively. (D-E) Word clouds of all topics in COVID-19 policy documents published before (D) and after (E) the WHO's pandemic declaration (March 11, 2020). Throughout the figures, the black dashed line marks the date of the WHO's pandemic declaration.

consistent with the idea that policy uses the science it cites. Meanwhile, we find that the science cited by policy documents tend to be of high quality, and policy documents that cite science also turn out to be highly cited within the policy domain, showing a nexus between impactful scientific research and policy work. The following sections document more details about these findings.

3.2. Quantifying COVID-19 policy landscape

3.2.1. Synchrony between pandemic and policy

As a first look at the policy data and its practical relevance, we examine how the evolution of COVID-19 policy documents corresponds to facts on the ground. To this end, we compare the share of COVID-19 policy attention and total confirmed deaths from COVID-19 over time. The policy documents mirror the case dynamics (Fig. 3.5A), showing a remarkable synchrony between the share of COVID-19 policy documents among all policy documents and the number of total confirmed cases.

We further compare the share of COVID-19 policy attention and total confirmed deaths from COVID-19 over time, finding a high degree of similarity between the two trajectories (Fig. 3.6A). To test if there are systematic delays between the share of COVID-19 policy attention and the pandemic progress (both cases and deaths), we further calculate the cross-correlation between the two time-series. To account for the approximate exponential growth of both curves, we take the first-order difference of logarithm transformation, defined as

$$Y_t \;=\; \log_{10} X_{t+1} - \log_{10} X_t, \tag{3.1}$$

where X_t is the original time series. Then, we calculate a normalized cross-correlation function. Specifically, given two transformed series Y_t and Z_t, the function is defined as

$$Corr(\Delta t) = PCC(Y_t\ , Z_{t+\Delta t}), \tag{3.2}$$

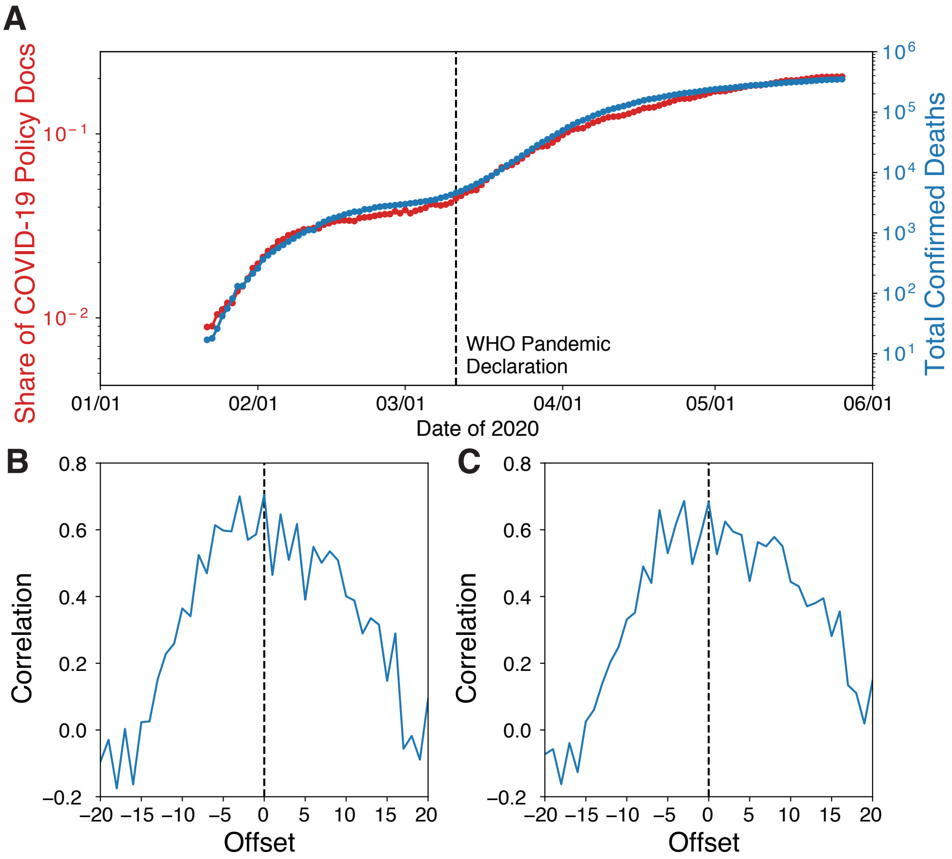

Figure 3.6. **Share of COVID-19 policy docs and the pandemic response.**

(A) Same as Fig. 3.5A, but for number of deaths. (B) Pearson correlation between the shifted COVID-19 share curve and the total cases curve. (C) Pearson correlation between the shifted COVID-19 share curve and the total deaths curve. The offset is positive when shifting the COVID-19 share curve forward.

where PCC is the Pearson correlation coefficient. The correlations between the transformed series suggest that the relationship between the time series is closest when the time offset, $\Delta t = 0$, between the share of COVID-19 policy documents and COVID-19 deaths is close to 0 (Fig. 3.6BC). In particular, when the offset $\Delta t = 0$, the cross-correlation is 0.705 for cases and 0.682 for deaths, respectively. These results suggest a high degree of synchronicity between the share of COVID-19 policy attention and how the pandemic evolves.

3.2.2. Field and topic evolution of COVID-19 policy documents

We further examine the content of the COVID-19 policy documents, by leveraging field and topic classifications from Overton and Dimensions to determine the primary focus of each policy document (Fig. 3.5). Overton uses machine learning approaches to assign fields and topics to policy documents. The Overton policy field classification is primarily based on the International Press Telecommunications Council (IPTC) Subject Codes taxonomy, the global standards body of the news media. In our analysis, we use 18 top-level fields in Overton and further group them into three major field categories:

Science & Health: “health”, “science and technology”.

Economy & Labour: “economy, business and finance”, “labour”, “prices”.

Society & Others: “arts, culture and entertainment”, “conflicts, war and peace”, “crime, law and justice”, “disaster, accident and emergency incident”, “education”, “environment”, “human interest”, “lifestyle and leisure”, “politics”, “religion and belief”, “society”, “sport”, “weather”.

We first group COVID-19 related policy documents into three major field categories and observe clear shifts in policy attention related to the pandemic (Fig. 3.5B). In the early stage of the outbreak (January and February 2020), about 90% of COVID-19 policies belong to the health and science category, showing a clear, initial focus on medical and public health issues. The policy priorities show a visible shift, however, since the WHO declared COVID-19 a pandemic on March 11th, 2020, with a notable rise in attention to issues around the economy and society, suggesting a growing policy balance between health and socio-economic implications of the pandemic.

The Overton policy categorization also implements a more fine-grained classification system. Here, we leverage this classification scheme to further examine the content of the COVID-19 policy documents. At the narrower topic level, here we show the top 10 topics by volume (Fig. 3.5C), after excluding one generic topic ("coronavirus disease 2019"). We find a clear decrease in the share of health-related topics and an increase of topics related to the policies, human activities and society since early March 2020. We also show the word clouds of topics pre- and post-pandemic declaration by the WHO respectively (Fig. 3.5DE), which further illustrates the evolution from public health toward social and economic issues.

We further test the robustness of these observations by looking into the data by month and for individual fields. We find similar shifts in topics of COVID-19 policy documents around March 2020 (Fig. 3.7A), with January and February primarily focusing on public health and medicine while April and May showing more expansive subject orientations and increased focus on social and economic issues. Further, we calculate the share of COVID-19 policy documents by each field and plot the results for the top 10 fields ranked by total COVID-19 policy documents (Fig. 3.7B), finding similar shifts in COVID-19 policy attention from health to economy and society.

Figure 3.7B features "politics" and "crime, law and justice" as the top two fields under our "society and others" category. We further break down the high-level category of "society" into more fine-grained levels (Fig. 3.7B inset). As Overton implements a

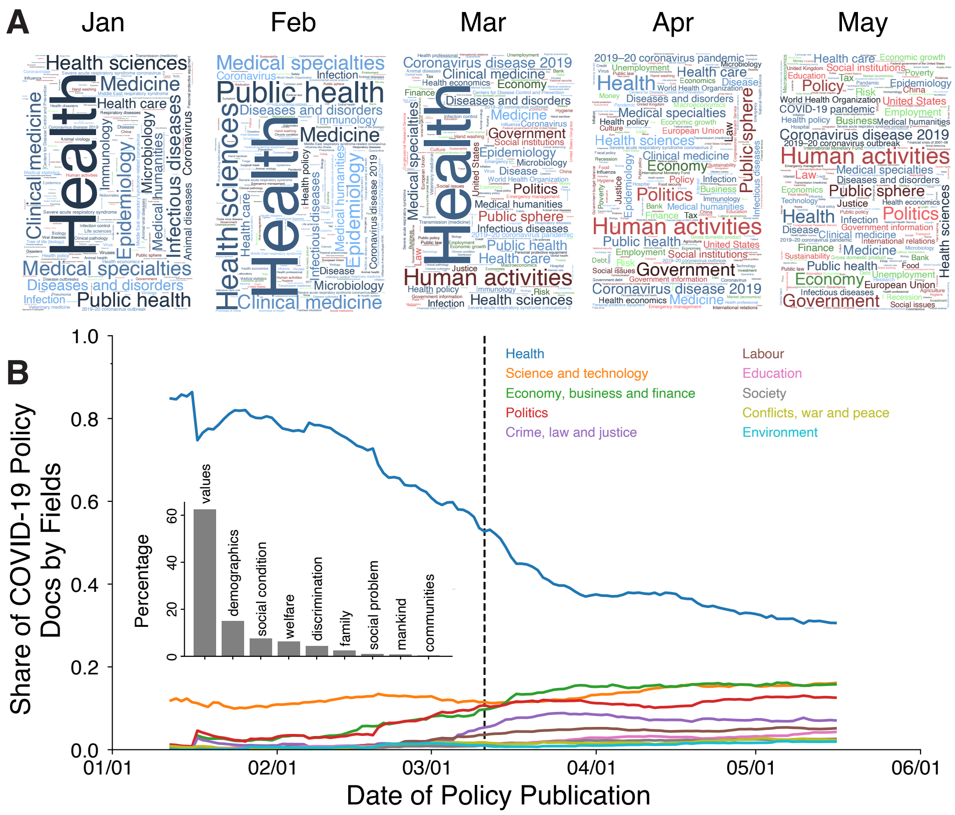

Figure 3.7. **Field evolution of COVID-19 policy documents.**

(A) Word cloud of topics, by month, of COVID-19 policy documents published in the first five months of 2020. The size of a topic corresponds to the share of the topic among all topics in the documents. (B) The share of total COVID-19 policy documents across time by fields (21-day moving average).

hierarchical classification system, we find that most society-related policy documents are classified into subcategories of “values” and “demographics”. Upon closer inspection of our data, we find that “values” subcategory includes specific discussions on death and ethics, which is relevant in the context of the COVID-19 pandemic. At the same time, many “demographics” policy documents are about immigration and international travel, which also represent an important policy focus in this pandemic.

Notably these shifts are observed in COVID-19 policy documents only. As we repeat our analyses on the evolution of fields for other (i.e., non-COVID-19) policy documents published in the same period. We find that the shares of policy documents by three major field categories (Fig. 3.8A) and individual fields (Fig. 3.8B) stay relatively stable over time, suggesting that the shifts are observed in COVID-19 policy documents only.

3.3. Quantifying uses of science in COVID-19 policy documents

Much like the global policy frontier, the scientific understanding of COVID-19 also evolved rapidly, as exemplified by the strong response from the global research enterprise. According to Dimensions data, more than 40,000 papers on coronavirus research were published from 1 January through 30 May 2020. Here, we uncover close connections between the evolving COVID-19 policy frontier and the evolving scientific frontier.

3.3.1. Close connection between science and policy

We calculate the fraction of COVID-19 policy documents that cite at least one scientific paper, finding it fluctuates in early 2020 but then features a steady increase with time, especially after the WHO's pandemic declaration (Fig. 3.9A). This observation indicates the close connections between the evolving COVID-19 policy frontier and the evolving scientific. In addition to a binary variable for whether a policy document cites science, we also alternatively consider the count of scientific references in a given policy document. Specifically, we plot the average number of scientific references in policy documents over time (Fig. 3.9C).

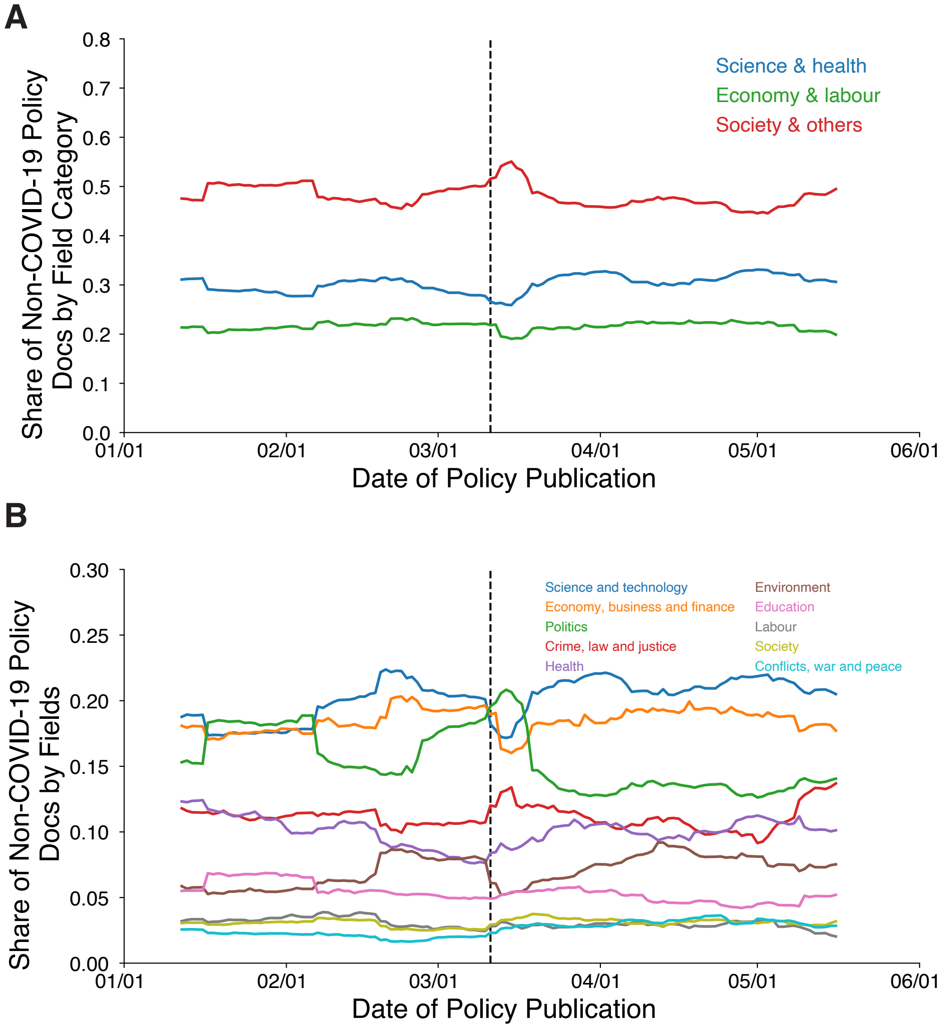

Figure 3.8. **Topic and field shifts of all other policy docs.**

(A) The share of total non-COVID-19 policy documents across time by three field categories (21-day moving average). (B) The share of total non-COVID-19 policy documents across time by fields (21-day moving average). Only the top 10 fields are presented.

Further, COVID-19 policies are disproportionately centered on the latest scientific frontier (Fig. 3.10A). Indeed, out of all scientific references drawn upon by COVID-19 policy documents, 19.9% of the scientific papers were published in 2020. This rate of utilizing the newest science is highly unusual, more than ten times larger than seen for

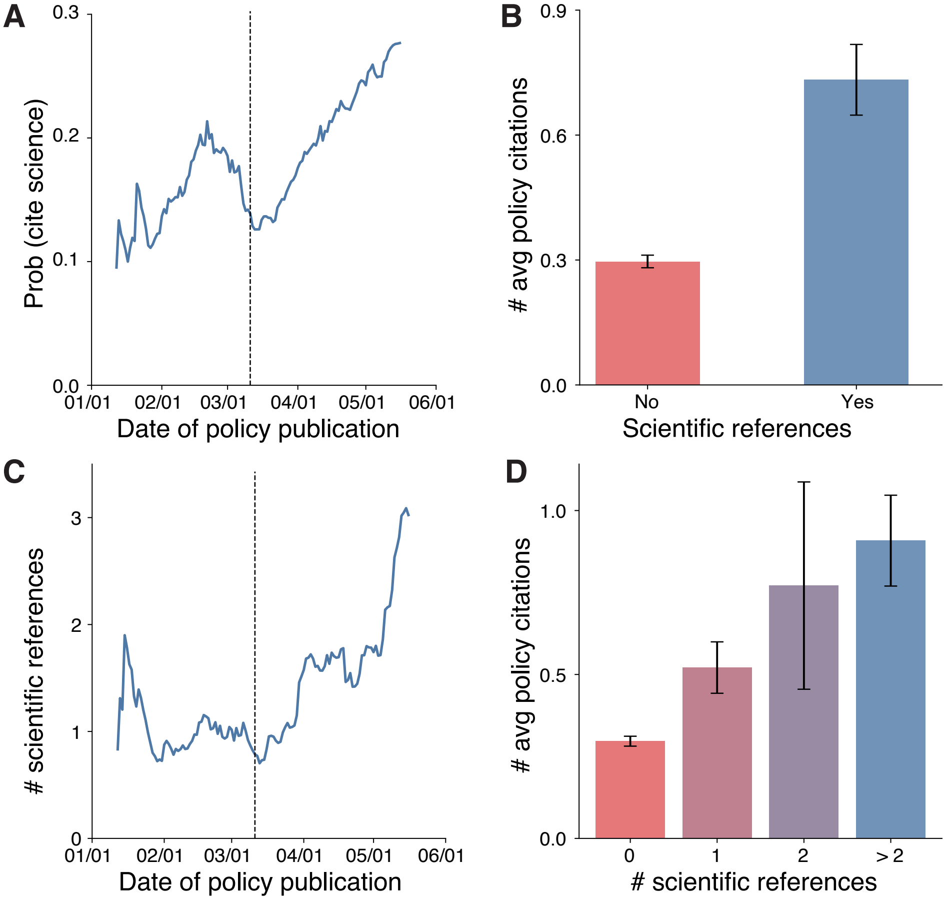

Figure 3.9. **Policy citations to science.**

(A) Probability of citing scientific references for COVID-19 policy documents published in 2020 (21-day moving average). The black dashed line marks the date of the WHO's pandemic declaration. (B) COVID-19 policy documents that cite scientific papers are much more likely to be cited by other COVID-19 policy documents. Error bars represent standard error of the mean. (C) Average number of science references cited for COVID-19 policy documents published in 2020 (21-day moving average). (D) Average policy citations as a function of number of scientific references cited. Here samples with more than 2 scientific references are combined so that each bin has comparable sample sizes. Error bars represent standard error of the mean.(A) The share of total non-COVID-19 policy documents across time by three field categories (21-day moving average). (B) The share of total non-COVID-19 policy documents across time by fields (21-day moving average). Only the top 10 fields are presented.

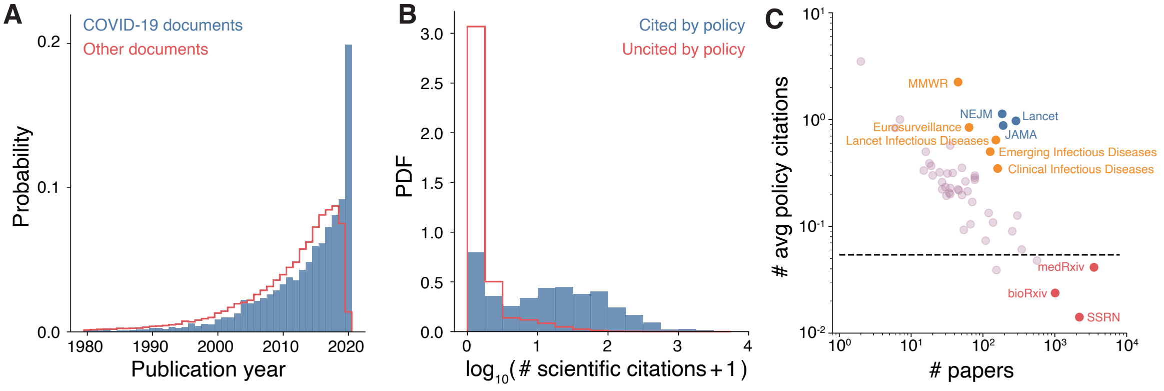

Figure 3.10. **Science use in policy documents.**

(A) Distribution of publication years of scientific papers (published from 1980 to 2020) cited by policy documents. The unusual spike in citing papers published in 2020 indicates that COVID-19 policy documents draw heavily on recent scientific evidence. (B) COVID-19 scientific papers that are cited by policy documents have greater citation impact within science. (C) For different journals and preprint servers, we measured the number of COVID-19 related papers (x axis) and the average number of citations from COVID-19 policy documents to these papers (y axis) in 2020. Shown here are the top 50 publication outlets based on the total number of citations from COVID-19 policy documents. The black dashed line indicates the average number of citations measured on all COVID-19 papers. MMWR, Morbidity and Mortality Weekly Report; NEJM, The New England Journal of Medicine; JAMA, The Journal of the American Medical Association

other policy documents. Not surprisingly, the latest science cited is primarily related to COVID-19 (88.4%). The close connection between science and policy is also reflected in the fields of science that COVID-19 policy documents cite, showing a clear shift from drawing primarily upon the biomedical literature to citing economics, society, and other fields of study, consistent with overall shifts in policy focus. Indeed, Dimensions implements the Fields of Research (FOR) classification for scientific papers. The FOR is a component of the Australian and New Zealand Standard Research Classification (ANZSRC) system, which follows a three-level hierarchy (divisions, groups and fields) and covers a broad set of research fields from the sciences and engineering, social sciences, and arts and humanities. In our analysis, we use 22 top-level divisions as the research fields of scientific papers. Analogously, as we do for policy documents, we further group 22 top-level divisions into the same three major field categories (Fig. 3.11A):

Science & Health: “Agricultural and Veterinary Sciences”, “Biological Sciences”, “Chemical Sciences”, “Earth Sciences”, “Engineering”, “Environmental Sciences”, “Information and Computing Sciences”, “Mathematical Sciences”, “Medical and Health Sciences”, “Physical Sciences”, “Psychology and Cognitive Sciences”, “Technology”.

Economy & Labour: “Commerce, Management, Tourism and Services”, “Economics”.

Society & Others: “Built Environment and Design”, “Education”, “History and Archaeology”, “Language, Communication and Culture”, “Law and Legal Studies”, “Philosophy and Religious Studies”, “Studies in Creative Arts and Writing”, “Studies in Human Society”.

Researchers have evaluated the reliability of field classification systems. For example, Ref. [129] empirically assessed the reliability of the Dimensions data. The Dimensions team subsequently followed up with a reply in Ref. [130], reporting further improvement based on more training data and better algorithms. Therefore, the current version of Dimensions data has two columns for the Field of Research (FOR) classification: “FOR” and “category_FOR.” Both columns are based on the same classification scheme but use different algorithms: “FOR”, as used in Ref. [129], is the more basic version, whereas

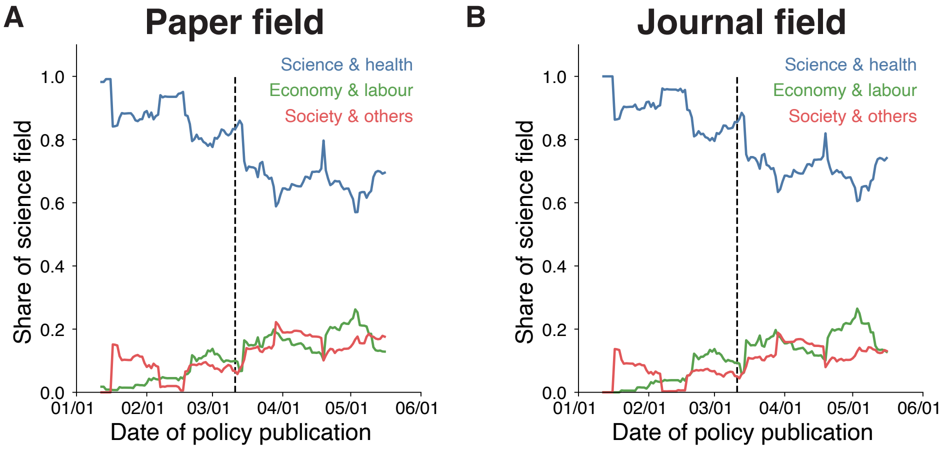

Figure 3.11. **Share of scientific papers cited by COVID-19 policy documents across three broad field categories (21-day moving average).**

We show the evolution using (A) the paper's field and (B) the journal's field. Throughout the figures, the black vertical dashed line marks the date of the WHO's pandemic declaration.

"category_FOR" is the improved version and the one that is officially recommended by Dimensions. In our analysis, we used "category_FOR" for field classification (i.e., the improved version).

As we further consider the share of scientific papers cited by COVID-19 policy documents for individual Fields of Research (category_FOR), we find a clear shift of COVID-19 policy documents from drawing primarily upon biomedical literature to citing economics, society, and other FOR. We further perform a set of robustness checks by calculating a major field of study for each venue (journal/preprint server) as an independent way of validating Dimension's field classification. This step, similar to the idea of assigning fields of study based on journals in Web of Science and Scopus, would reduce errors at the individual paper level and offer a more stable characterization. We find that results under the new field assignment (Fig. 3.11B) are consistent with the results in Fig. 3.11A, further documenting the robustness of our results.

Together, these results suggest that despite the extremely recent development in COVID-19 related research, new scientific work has rapidly found its way into policy documents.

3.3.2. Scientific impact and policy impact

The close relationship between science and policy prompts us to examine the quality of scientific evidence that informs policy.

Here we examine the quality of science appears in policy documents along two dimensions. First, we separate COVID-19 related papers into two groups based on whether or not they are referenced by COVID-19 policy documents, and measure each paper's scientific impact within the science community, approximated by the number of citations the paper receives from other scientific papers. We find a dramatic difference between the two groups (Fig. 3.10B): papers referenced in policy documents garner on average 40 times higher citations than those not referenced in policy (average citations: 67.72 vs 1.67).

Note that, citations are known to be dynamic over time [131]. Hence, when comparing or evaluating citations of papers, one should always take into account their publication year, as older papers tend to collect more citations. As we compared the citation distribution of COVID-19 papers, which are all published in 2020, grouping them by whether they have been cited by at least one policy document. Therefore, given the standard practice in citation analysis -- normalizing citation counts based on publication year – we

[]

Variables	Dependent variable $\log_{10}$(Scientific citations + 1)	
	(1)	(2)
D_cited_policy	0.957***	0.949***
	(0.012)	(0.012)
Paper date		-0.00045***
		(0.000033)
Constant	0.149***	0.196***
	(0.0018)	(0.0039)
Obs.	40732	40732
Adj. R2	0.144	0.148
F statistic	6848.4	3533.6

Table 3.1. **Regressions after controlling for publication date.**

The difference in citations counts from other scientific papers, comparing papers that are cited by policy documents with papers that are not. Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

did not control for publication time in Fig. 3.10B. Here we go one step further and run an additional regression that controls for the publication date (i.e., accounting for within year differences) of a paper. We find that the difference in citation distribution between the two groups remains large and significant (Table 3.1). Overall, this result shows that the coronavirus research used by policymakers aligns with what scientists heavily engage themselves.

Further, we break down the policy coverage of COVID-19 research based on publication venues (Fig. 3.10C). We find that different venues differ widely in publication volume, with preprint servers such as medRxiv, bioRxiv, and SSRN publishing an order of magnitude more COVID-19 related papers than peer-reviewed journals. Yet, despite the volume of preprints, their impact in policy is rather limited, as these preprint servers show consistently fewer policy citations than average. By contrast, COVID-19 policy documents disproportionately reference peer-reviewed insights, drawing especially heavily on top medical journals, both general (e.g., Lancet) and specialized (e.g., Clinical Infectious Diseases). Though peer review does not necessarily guarantee high-quality science [111], amid growing concerns over the quality and abundancy of coronavirus research posted on preprint servers, these results nevertheless show that during this crisis, peer-reviewed journals continue to remain a crucial institution in supplying scientific evidence for policy making.

We also approximate scientific “quality” using a citation-independent measure. Here we focus on COVID-19 related papers published on preprint servers, testing the possible relationship between being cited by policy documents and getting published in peer-reviewed journals. One technical challenge here is that it is generally not easy to link preprint publications to their peer-reviewed versions, given that papers may undergo significant changes during the review process. To overcome this challenge, here we leverage a novel data source from the NIH's Preprint Pilot database. We downloaded paper information for all preprint publications from multiple sources (medRxiv, bioRxiv, ChemRxiv, SSRN and Research Square) in NIH's iSearch COVID-19 Portfolio. One unique advantage of this dataset is its linkage between the preprint and published version of the same paper. After merging with our dataset of COVID-19 papers, we are left with 5,993 preprints, among which 1,344 papers (22.4%) can be linked to their peer-reviewed versions.

Our first analysis directly compares the number of policy citations between the two groups, finding that those preprint papers that have been published in peer-reviewed journals receive 2.7 times as many policy citations as those that have not (0.0558 vs 0.0204). We further run linear and negative binomial regressions that control for publication date and include preprint server fixed effects, finding the advantage of published papers remains

Model	(1)	(2)		(3)	(4)
VARIABLES			VARIABLES		
D_published	0.0354*** (0.0083)	0.0237*** (0.0084)	D_published	1.005*** (0.259)	0.524** (0.251)
Paper date		-0.00133*** (0.00014)	Paper date		-0.0431*** (0.0050)
[] Fixed effect for preprint server		Yes	Fixed effect for preprint server		Yes
Intercept	0.0204*** (0.0039)	0.145*** (0.024)	Intercept	-3.891*** (0.142)	-19.688 (9484.75)
Obs.	5993	5993	Obs.	5993	5993
Adj. R2	0.003	0.021	Pseudo R2	0.012	0.121
F Statistic	18.26	22.80	loglikelihood	-628.34	-558.84

Table 3.2. **Regressions comparing among COVID-19 related preprints.**

Here we consider COVID-19 related preprints, comparing policy citations to these papers depending on whether the preprint has been published in a peer-reviewed journal. Columns (1-2) are based on an ordinary least squares (OLS) model, while columns (3-4) are based on a negative-binomial model. Standard errors in parentheses. $^*P < 0.1$; $^{**}P < 0.05$; $^{***}P < 0.01$.

robust (Table 3.2). Note that, these differences likely reflect a conservative estimate of the overall difference in policy impact between the two groups, as preprints that have not passed peer review may be published at a later date, suggesting the counterfactual group (i.e., those that never pass peer review) may have an even lower policy impact. Overall, these results further strengthen the findings that policy documents are more likely to be grounded in peer-reviewed science.

Overall, the COVID-19 policy frontier appears deeply grounded in extremely recent, peer-reviewed scientific insights, and science directly drawn upon by this policy frontier appears especially impactful within the research community itself. Moreover, policy documents that are grounded in the scientific frontier also tend to garner substantially more citations within the global policy network. Specifically, separating COVID-19 policy documents by whether they cite science or not, we find that COVID-19 policy with policies that reference science receiving 0.436 additional policy citations on average, more than doubling the baseline rate (Fig. 3.9). As a robustness check, we further calculate average number of policy citations received by a policy document as a function of scientific references cited (Fig. 3.9D). We find a positive relationship between science used and policy citation impact, further supporting our main conclusions.

To test if this difference in use can be explained by other covariates, we further test the robustness of this result using a linear regression model (Table 3.3) that controls for the publication date of the document as well as fixed effects for the country and type of institution. We find that all else being equal, COVID-19 policy documents that cite scientific papers are associated with 0.322 more policy citations than those that do not ($P < 0.001$). We also separate the citations they receive into citations from the same institution or different policy institutions, finding additional citations for both types (0.184 more policy citations from the same institution and 0.137 more policy citations from the other institutions). We repeat the above regression analysis using a negative-binomial regression model (Table 3.4), finding our results remain largely robust.

Together, these results show that, despite the rapidly evolving nature of the pandemic, the policy and scientific frontier of COVID-19 are closely interlinked, with documents and articles directly along the policy-science interface (i.e., policy documents that cite science and the cited science itself) being notably more impactful within their own domains.

[]

Model	(1)	(2)	(3)	(4)	(5)
		total		same inst	different inst
VARIABLES					
D_cites_science	0.436***	0.516***	0.322***	0.184***	0.137***
	(0.053)	(0.052)	(0.054)	(0.042)	(0.023)
Policy date		-0.010***	-0.011***	-0.007***	-0.004***
		(0.001)	(0.001)	(0.001)	(0.000)
Inst type			Yes	Yes	Yes
Inst country			Yes	Yes	Yes
Constant	0.297***	1.302***	1.097*	0.732*	0.365
	(0.240)	(0.078)	0.585	(0.431)	(0.250)
Obs.	7730	7730	7730	7730	7731
Adj. R2	0.009	0.032	0.065	0.045	0.055
F statistic	68.84	125.61	9.48	6.83	8.09

Table 3.3. **The ordinary least squares (OLS) regressions including the policy document publication date, country, and type of institution.**

Columns (4-5) show the results using citations from the same/different institutional sources only. Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

[]

Model	(1)	(2)	(3)	(4)	(5)
		total		same inst	different inst
VARIABLES					
D_cites_science	0.905***	0.742***	0.43***	0.320***	0.517***
	(0.095)	(0.091)	(0.091)	(0.103)	(0.013)
Policy date		-0.025***	-0.026***	-0.024***	-0.028***
		(0.002)	(0.002)	(0.002)	(0.002)
Inst type			Yes	Yes	Yes
Inst country			Yes	Yes	Yes
Constant	-1.216***	1.113***	-19.516	-19.769	-16.781
	(0.046)	(0.161)	(16925.42)	(18910.6)	(4040.64)
Obs.	7730	7730	7730	7730	7730
Pseudo R2	0.010	0.036	0.092	0.083	0.137
loglikelihood	-5016.59	-4884.06	-4602.05	-3726.35	-1977.36

Table 3.4. **The negative-binomial regressions including the policy document publication date, country, and type of institution.**

Columns (4-5) show the results using citations from the same/different institutional sources only. Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

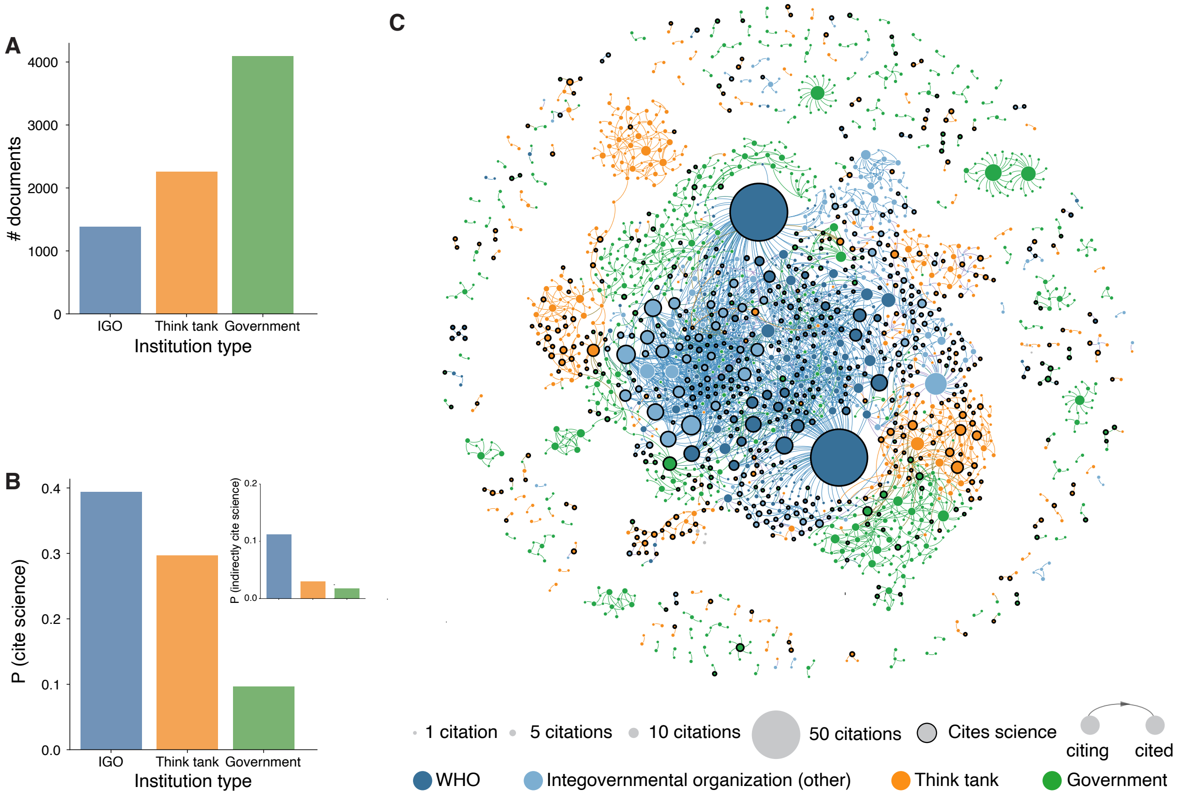

Figure 3.12. **The COVID-19 policy citation network.**

(A) Number of COVID-19 policy documents published by institution type. (B) Probability of citing science by institution type. Inset: Probability of indirectly citing science by institution type (citing other COVID-19 policy documents that in turn cite science). (C) Network visualization for the COVID-19 policy document citation network. Each node corresponds to a COVID19 policy document, colored by the institution type to which it belongs. For visualization purposes, only nodes with at least one link are shown in this network. A link between documents is colored by a mixture between colors of the source and target nodes. The size of each node is proportional to the number of citations it receives from other COVID-19 policy documents. A node border (in black) indicates policy documents that draw on scientific papers.

3.3.3. The role of policy institutions

What policy institutions contribute most strongly to the policy-science interface? Our final analysis visualizes the policy citation network among COVID-19 policy documents (Fig. 3.12). This network conveys three key insights. First, hubs in this policy network disproportionately cite science. Second, although government agencies produced the most COVID-19 policy documents among the three types of institutions Fig. 3.12A), they are the least likely to cite science (Fig. 3.12B), and their positions in the network are largely peripheral, sometimes even separated from the main cluster.

By contrast, policies that are central to the network, and especially those grounded in science, are disproportionately produced by IGOs, especially by the WHO (Fig. 3.12BC). These differences in the use of science persist when we compare the indirect use of science (i.e., citing other policy documents that cite science), showing that IGOs again draw disproportionately more on the policy-science interface (Fig. 3.12B inset). Many have argued that nations work best together through international institutions, especially in a crisis like COVID-19 [132]. These results suggest a key role of the WHO and other IGOs in the global policy response to COVID-19, acting as central conduits that link policy to science.

3.4. Concluding Remarks

Taken together, our results show that policy documents in the COVID-19 pandemic substantially access recent, peer-reviewed, and high-impact science. At the same time, our reference-based measures are but a proxy for the uses of science in policy [103] and policies may cite science for different reasons [108]. Policy-relevant science may be interpreted differently depending on one's specific interests [106] and may even be distorted during the dissemination process [107]. Further, although our data captures among the largest collection of policy documents, there could be potential biases in data sample and coverage that future research may help to further elucidate. Also, our data capture science-policy interactions up to May 26th, 2020, and the observed patterns may continue to evolve as the pandemic unfolds worldwide. Nevertheless, our results suggest that COVID-19 policy documents appear neither isolated from scientific advances nor reliant on dubious science. These findings appear encouraging for the scientific community, as scientists, journals, and funders work expeditiously to advance and validate new research, with the hope that their work might impact the course of the pandemic.

Ultimately, although scientific advances provide a global public good, and IGOs can help coordinate global action, national policy approaches and death rates have varied dramatically [133]. While some countries have been quite successful in containing the outbreak [134], some have been actively antagonistic to IGOs and scientific advice [132, 135]. In the current picture, science is breaking through, and scientific results are being heard, but they are not being heard everywhere.

CHAPTER 4

Science as a public good: Public use and public funding of science

Science is often seen to provide substantial impacts beyond the community of scientists themselves — for technological progress, government function, basic human curiosity, and more [1, 2, 136–142]. Given the potential benefits, many nations have built institutional architectures to support science through public investment, following the logic of public goods [143–145]. Like a public park, which is funded by the government and can be visited for free, scientific research is substantially funded by governments with its results placed in the public domain. This institutional design seeks to enable broad use of scientific ideas and avoid underinvestment by private actors. Yet in turning to public funding, this approach relies on the idea that public investment in science can match the public interest in science.

Although public investment in science is a central feature of the scientific ecosystem [144–146], empirically examining the varied public uses of science and testing whether there is alignment between public funding and public use has remained elusive, mainly due to the difficulty in collecting systematic data. Moreover, the lack of measurement has invited substantial skepticism. Indeed, many observers view scientific research as a cloistered or 'ivory tower' activity that rarely corresponds to the public interest [147–151]. For example, the "two communities" and "two cultures" theories highlight substantial knowledge and interest gaps between scientists and policymakers, disconnecting scientific research from policy insights [103, 109, 152, 153] and suggesting little relationship between the quality of research and its public usage [104, 152, 154]. Meanwhile, scientists may have peculiar interests, with little exposure to real world problems or incentives to tackle them [141, 155]. These potential gaps further animate root concerns over the public funding of science and its proper allocation [156–159]. For example, policymakers have long criticized the National Science Foundation (NSF) for funding frivolous research and have called for greater transparency around the relevance of science [156, 157]. Some prominent academics and commentators, including Nobel-Prize winner Milton Friedman, have taken the position that the government should not fund science, favoring purely private sector research instead [158, 159].

4.1. Related works

4.1.1. Policy uses of science

There is long-standing interest in understanding uses of science in policy and decision-making over many decades [103, 104], offering a rich set of conceptual models capturing policy uses of science, ranging from (a) two communities theory [109, 152] that suggest little common language and interest between scientists and policymakers, to (b) supply-side and demand-side models that highlight knowledge production on the science side [160] or problem solving on the policy side [108] as the primary driver of research utilization in policy, to (c) interaction models that depicts an iterative process where science and policy co-evolve [103, 161, 162]. Some works along this line have also examined different types of policy uses, arguing that policy-science citations may occur for various reasons

[106, 107], including (i) instrumental uses (knowledge directly applied to solve problems); (ii) conceptual uses (research that influences or informs the way policymakers think); (iii) tactical uses (citing research to support or challenge an idea), among others, suggesting value in understanding the semantics of the policy-science citations. Yet at the same time, it has also been recognized that distinguishing these uses at scale remains a challenging task [103]. Here we develop a scaled approach based on citation links and leave analysis of how specifically the science is being used to further investigation.

4.1.2. Altmetrics literature

Altmetrics studies alternative or complementary indicators related to scientific publications [126, 163, 164] . This field leverages databases such as Altmetric and PlumX and has grown rapidly in the last decade, deepening our empirical understanding [126] of how science is covered across different online platforms, including social media platforms, mainstream media news, policy documents, Wikipedia and other sources. For example, several prior studies in this literature have examined the overall coverage and reliability of such datasets [165–167]. Based on these datasets, researchers have also examined how scientific papers from different fields receive attention beyond citations [168], as well as the relationship between citation metrics and altmetrics indexes [163, 164]. For a more detailed overview of the Altmetrics literature, readers may refer to two recent reviews [126, 169].

4.1.3. Scientific non-patent references

Scientific non-patent references have also been studied in the recent innovation literature, partly due to the increasing availability of approaches and data sources for large-scale matching between patent references and scientific publications. Recent systematic linkage efforts have connected USPTO patents to Web of Science [170] and Microsoft Academic Graph [171]. Existing literature has examined meanings of such citations, suggesting these linkages as a useful signal for association between science and technology [172, 173]. Furthermore, scientific non-patent references have been leveraged to construct higher-order indirect links between patenting and science [174, 175].

4.1.4. Science as a public good

The outputs of scientific research are often described as a form of public good [107, 145]. In economics, public goods are formally characterized by two properties: non-rivalry and non-excludability. Non-rivalry means that a good's use by one party does not constrain its use by another party. For example, the use of algebra, a machine learning algorithm, or polymerase chain reaction by one party does not make it unavailable to another party. Indeed, a remarkable thing about ideas is that the same idea can be used by large numbers of people at the same time. This feature points to the potential wide social benefits from the creation of new ideas. It also stands in contrast to more ordinary properties of "rival" goods (e.g., a chair, a computer, a car) where the use by one party prevents the simultaneous use by another party.

The second feature of public goods is non-excludability. Excludability refers to whether one can prevent others from using the good. Note that, whereas non-rivalry is an innate property of ideas and the outputs of research, non-excludability is in part a policy choice. For example, patent law allows inventors to take a new idea and turn it into a private good (the patent allows the inventor to exclude others from using it) rather than a public good. Treating scientific research as a public good is then in part a policy choice. That is, the science system usually seeks to make new insights widely available. Other examples of public goods include the national defense or a public park, which can be enjoyed by many people at once and where excluding certain groups from the national defense or a public park is either difficult or non-desirable.

A key implication of public goods is that markets will underprovide them. In particular, due to non-excludability, consumers can make use of a public good without compensating the creator. Since the creation of a public good (a research result, a public park, the national defense) is costly and the investor will have difficulty recovering these costs from users, the private incentive to invest in public goods is weak. With markets under-providing public goods, we therefore turn to public policy to support investment in public goods. In practice, governments often cover the investment costs by investing directly in public goods. For example, governments invest up-front in scientific research projects (rather than seeking ex-post compensation as in the patent system) and makes the insights from these projects publicly available, thus embracing the non-rival nature of ideas and extending access widely to maximize the benefits of new insights. In stepping away from a market mechanism, however, this institutional approach presumes that the scientific apparatus, and the public investments in it, are in fact aligned with public use. For a detailed discussion of the properties of public goods, with applications to the under-provision of scientific research, see sources such as [107, 145, 176].

This paper advances a measurement framework to study public uses of science, the public funding of science, and how public use and public funding relate. Building on prior research that considers the use of science within a given public domain [163, 164, 171, 174, 175, 177], here we integrate five large-scale datasets that link scientific publications from all scientific fields to their upstream funding support and downstream public uses across three public domains.

4.2. Data description

4.2.1. Microsoft Academic Graph

The publication and citation data are primarily obtained from Microsoft Academic Graph (MAG, accessed Oct 2018) [178, 179]. MAG is among the largest open-source citation databases to date and contains records of 209 million documents. We inter-linked different data tables to obtain the author, affiliation, year, publication venue and field information for each paper. MAG includes a variety of document types. To focus on scientific articles, we consider publications under the categories of journal papers, conferences papers, books and book chapters, and papers with DOI information. In other words, two kinds of publications are excluded in our analysis: patents and papers with neither category nor DOI information. We further focus on papers published in a 10-year period from 2005 to 2014, leading to a subset of 36 million papers in total.

MAG uses a non-mutually exclusive hierarchy for research topic (field of study) mined by semantic analysis tools. To explore major research fields, the analysis considers the level-0 (19 fields) and level-1 (294 fields) categorizations (Fig. 4.1a). Figure 4.1b and 4.1c

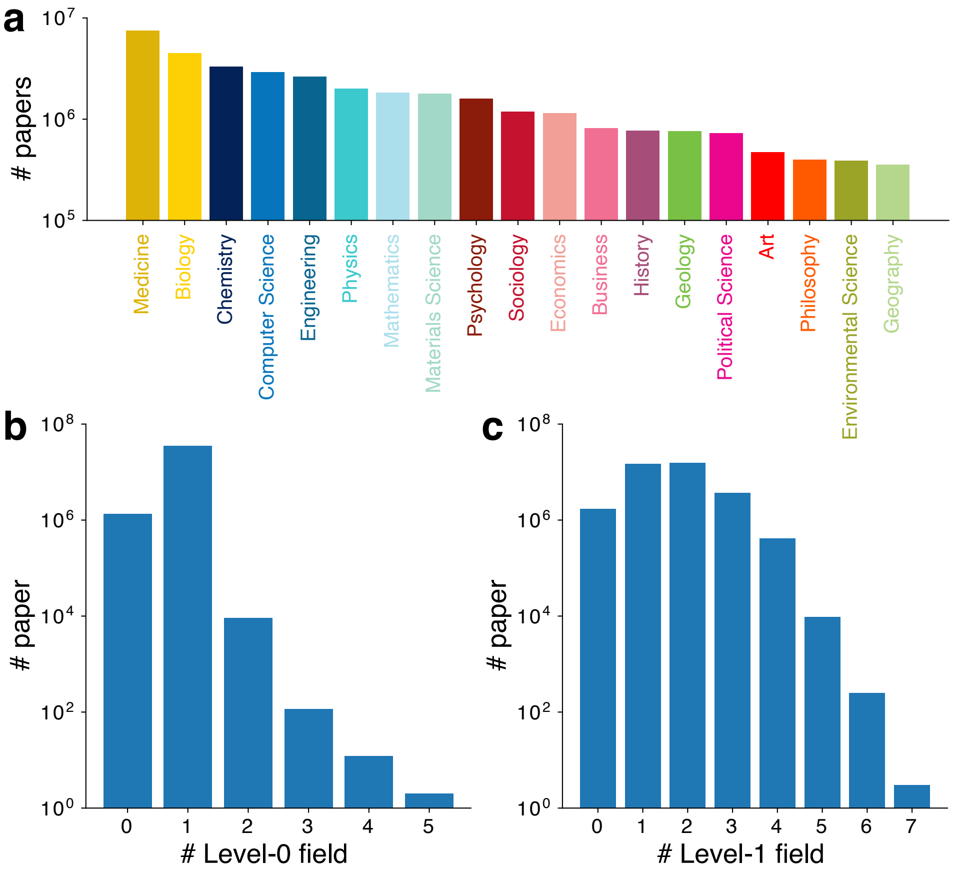

Figure 4.1. **An overview of MAG field classification.**

(a) Number of papers belonging to each of the 19 level-0 fields. **(b-c)** Distribution of field numbers that connect to a paper at level 0 (b) and 1 (c).

show the number of level-0 and level-1 fields that each paper is connected to, indicating both quantities are narrowly distributed.

4.2.2. US Government Documents

To quantify references to scientific articles in the government domain, one needs to construct a large-scale dataset of government documents that can be linked to the scientific papers. The task has been difficult in part because government documents are spread across many sources. Despite recent efforts like govinfo [180] to digitalize and standardize government publication information, most existing sources have relatively low coverage, especially for the executive branch. Furthermore, although a significant fraction of such documents may cite scientific literature, such citations do not follow a common structure. To tackle both challenges, here we develop a novel pipeline to construct the dataset.

Our data collection starts with a list of URLs under the .gov domain, which is the domain name for government agencies and contains the vast majority of U.S. government entities. Given the huge number of such pages, here we use a PageRank-like assessment system provided by the Microsoft Bing search engine, which assigns each URL with a tiered index of importance. In this study, we focus on Tier 0 (the most important) pages to construct the sample, which contains approximately 6 million URLs within the .gov domain. We downloaded these pages using an automatic crawler and focused on all PDF files in this set (˜28% among the corpus). We also notice that a small proportion of these documents are themselves research papers, as their urls can be linked to MAG papers though MAG paper url table, and exclude these documents from our analysis.

To extract the references cited in these files we use Science-Parse [181], an open-source tool for reference string extraction developed by the Allen Institute for Artificial Intelligence. Science-Parse is a state-of-art framework that scans PDF files and returns a list of all reference-like strings. We then matched this list to the MAG. Since the PDF reference extraction may contain minor errors, exact title matching of paper items may not be the optimal approach. Specifically, we indexed the full MAG to compile a search engine-like system using title, journal, author, and publication year information.

By leveraging the Okapi BM25 measure [182], one of the core algorithms used by modern full-text search engines, we query each string to obtain a list of the top 2 candidate paper items, each accompanied by a score representing the degree of agreement. To find the score threshold for determining if a string is successfully matched, here we use scores of the 2^{nd} matched paper as a null model for score distributions (Fig. 4.2a). Indeed, assuming the score is a reasonable quantification, one would expect the difference of 1^{st} and 2^{nd} matched paper to be significant if and only if the 1^{st} ranked paper is a true match of the string. To this end, for each string we first calculate the score distribution of all 2^{nd} matched scores of the similar query word length as a baseline. The string is considered to be matched to the 1^{st} ranked paper when the score is significantly higher than a right tail cutoff of the baseline distribution (one-sided $P = 0.05$, or equivalently, $Z = 1.65$). We further test this algorithm by comparing its predictions with manual validations on 100 randomly selected papers through two evaluations: (1) For a binary classification based on whether a reference string can be matched to a MAG paper, we calculate the F1 score; and (2) conditional on being classified as positive in (1), we measure the accuracy of the matched MAG document ID. We find high consistency between the results returned by the algorithm and our manual validations (Fig. 4.2b). See also Chapter 3 for additional validation analysis using the Overton data.

4.2.3. Altmetric dataset

To study references to scientific publications in the news media, we use a dataset offered by Altmetric [163, 164, 183]. This dataset records approximately 26.2 million papers with at least one news media or social media mention. We then merge paper information with

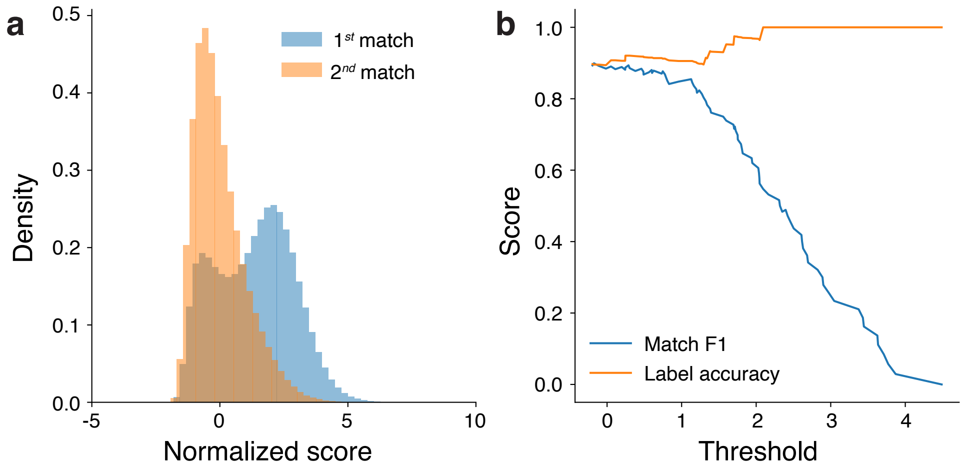

Figure 4.2. **Matching raw reference strings to MAG.**

(a) The distribution of normalized score for papers with the first and second highest matching score. The normalization is obtained by calculating the z-score of the raw score for the second-best matched papers for strings of similar word length. The normalized score for second-best matched papers approximately follows a standard Gaussian distribution, yet that for first-best matched papers show another mode that is larger than $Z = 1.65$, indicating a large proportion of matchings are significantly more accurate than expected. **(b)** We tuned the threshold and evaluated matching performance on a manually validated subset using two step measures: (1) Whether a string can be matched into a MAG paper (binary classification problem), measured by F1 score, and (2) Conditional on the string successfully matching into MAG both automatically and manually, to what extent are the two matched IDs consistent (label problem), measured by overall accuracy.

MAG. A vast majority (22.1M million) of such publications in the Altmetric database have unique digital object identifiers (DOI), allowing us to connect this with DOI information in MAG. We find that 17.2 million (78%) of the DOIs can be matched to records in MAG.

4.2.4. USPTO patent database

To study references to scientific publications in patents, we build on prior work and use a high-scale mapping from United States Patent and Trademark Office (USPTO) patents to

MAG papers, which includes approximately 31.7 million citation pairs between patents and papers [171, 184], from both the front page and full text of the patents. To classify patents into technology classes, we use the Cooperative Patent Classification (CPC) system, drawn from PatentsView, a data platform based on USPTO bulk data [185]. Combining the two files provides technology class information for 97.5% of patents that reference scientific articles. The small share of missing technology class cases corresponds to patents recently granted, which have not been updated in our data.

4.2.5. Dimensions scientific funding data

To understand how research funding from various sources is allocated into different scientific fields, we leverage research funding data from Dimensions [116, 186], which includes approximately 5 million research projects supported by over 400 funding agencies worldwide. To be consistent with our publication analysis, we focus on projects funded during the same ten-year period (2005-2014). One challenge in our estimation here is that some projects are supported both within and outside the ten-year period (for example, in years 2014, 2015 and 2016), while only the total funding amount is available. Here we estimate funding amount in the ten-period by multiplying the total amount with the fraction of time within the ten-year period, which equals to 1/3 in this example. We further focus on projects that have funding amount information and are funded by US agencies.

A unique opportunity provided by Dimensions is a linkage table between supporting grants and resulting publications, which allows us to categorize the field of each grant according to its resulting publications. More than 90% of the publications have DOI information which can be further matched to items in our paper database. Here we use this table and focus on projects that can be linked to at least one MAG paper (which allows us to estimate the research field for over 74% of U.S. research funding in this period). We then create a list of level-1 fields by combining the fields of resulting publications supported by each grant, and evenly split the funding amount of a project to each field in this list. For example, if a grant of $15,000 supports three publications, two in quantum physics and one in mathematical physics respectively, we assign $10,000 to quantum physics and $5,000 to mathematical physics. Together we link 292,875 funded projects with at least one publication.

4.2.6. Data limitations

Our data are not without limitations. First, our datasets only represent a subset of all possible government documents and media news in the world, and there could be heterogeneity within documents published by different agencies or news covered by different media. Second, the linkage strategy between science and public uses and funding is based on automatic algorithms and may contain some errors. While our validations in Supplementary Note 2 and robustness checks in Supplementary Note 5 have not uncovered any potential biases, readers should keep in mind of the existence of these factors. In addition, our analysis is primarily focused on PDF documents. The focus on PDF files is consistent with common practices in commercial products such as Altmetrics, partly due to the fact that PDF documents are more likely to cite scientific literatures. Extracting scientific references from unstructured documents at scale has remained a significant technical challenge. The machine reading and reference extraction technologies for PDF documents we developed in this paper, adapted from the pipeline for constructing Microsoft Academic Graph, are among the state-of-the-art in their kind. To the best of our knowledge, there is no similar approach to extract scientific references from HTML pages.

Nevertheless, despite these potential limitations, it is important to note that these data sources are among the largest in their respective domains, and approaches for data linkages are also among the most advanced of their kind, hence representing the state-of-art empirical basis to understand the interaction between science and public domains. For government documents, we further conduct additional validation exercises using the Overton data, a large-scale database of policy documents (see Supplementary Note 2.1 below). As another kind of validation analysis, we also observe the RCI at the sub-agency level and find for example that the U.S. Department of Treasury draws especially on economics and business research, while the U.S. Department of Energy draws especially on geology and engineering. Examining the sub-agencies produces rich patterns that appear to have substantial face validity, where the scientific areas drawn upon are closely related to the agencies' missions, and further suggests the substance of the linkages our analysis uncovers.

There are also additional channels of knowledge flows beyond those we trace through references. Collecting large-scale empirical records of these alternative channels remains a challenging task – especially given the fact that large-scale empirical data on the industrial use of science and social science is often harder to assemble than governmental and media use in the public space. As one example within our data, we further leverage the media use in our D_3 and focus on a general management magazine – *Harvard Business Review* (HBR). By tracking all citations from HBR articles to scientific publications we calculate

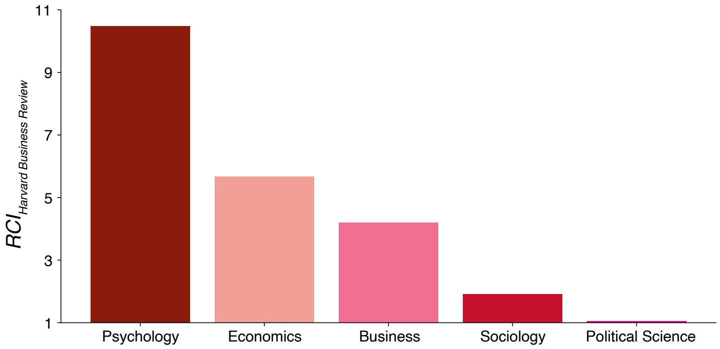

Figure 4.3. **Relative consumption index in Harvard Business Review articles for five L0 fields.**

the relative consumption index for this management-oriented outlet. We find that the top *RCI* fields in HBR are psychology, economic, business, sociology, and political science (Fig. 4.3), indicating management orientations that extend private sector interests beyond the research areas that are prevalent in the patenting sphere. More generally, addressing additional avenues will require new data approaches that go beyond our orientation on reference linkages in publicly available data.

4.3. Independent data validation

4.3.1. Overton policy documents

Policy documents extracted from the Bing search engine and their associated references provides a novel dataset that is only possible with recent advances in information retrieval and machine learning. Given the novelty of such applications (reference parsing in policy documents), we lack systematic baseline methods for comparison. Here we leverage another novel dataset, Overton, which has just become available during the writing of this manuscript and provides an independent validation case. Overton is among the world's largest searchable index of policy documents, including about 3M policy-related documents from thousands of sources (including government agencies, think tanks and intergovernmental organizations). Overton also extracts scientific references in documents and maps them into DOIs. Here we retrieve all policy documents published by U.S. governmental agencies indexed by Overton, looking at all scientific references (published in the same ten-year period) they have ever cited, and use the DOI to connect these papers into MAG. Chapter 3 documents the consistency between the two datasets. We further use Overton data to repeat our main results (see Appendix for details).

4.3.2. RePORTER funding dataset

The Dimensions data (D_5) is a state-of-art database linking funding and associated publications, using information from funding agencies as well as text mining from the acknowledgement section of publications. As an alternative, we further leverage funding-paper linkage information from two major funding sources in the U.S. – the National Science Foundation (NSF) and National Institutes of Health (NIH). These two agencies are the largest federal funders for scientific research and together account for more than half of overall federal research funding [187]. Information on all projects funded by NIH over the last several decades and resulting publications are available through NIH RePORT (Research Portfolio Online Reporting Tools), an open data source developed since 2008 [188]. Bulk data for NSF grants in a similar format are available as part of Federal RePORTER, a federal effort to "create a repository of data and tools that will be useful to assess the impact of federal R&D investments" [189].

To test our data coverage, we calculated the number of publications supported by each grant active in the ten-year period (2005-2014) from both RePORTER and Dimensions data. We find in both NSF and NIH, the number of publications supported by each grant is highly correlated, showing a high degree of consistency between the two data sources (Fig. 4.4bd). Further, we find that Dimensions reports more resulting publications on average (fig. 4.4ac). The superiority of Dimensions may be explained by multiple reasons, including an incomplete coverage of data in early years from Federal RePORTER and the fact that more complete paper-grant linkages can be found through publication acknowledgements. Regardless, these results suggest Dimensions is the preferred source for linking papers and grants for this study.

4.4. Empirical results

Our main results focus on papers published between 2005 and 2014, a common period covered by all three datasets, resulting in 128,465, 275,536, and 1,296,922 papers cited in government, news, and patent documents, respectively.

4.4.1. Diversity in public use

Our first analyses measure the usage of scientific research in the three public domains. To conduct this analysis, we first leverage the MAG's classification of papers across 19 top-level fields. To account for cross-field differences in publication volume, we define a Relative Consumption Index, *RCI*. For a given public domain (d) and field (f), *RCI* measures the fraction of papers in the field consumed by that public domain, normalized by the same fraction calculated on all fields for that domain. That is,

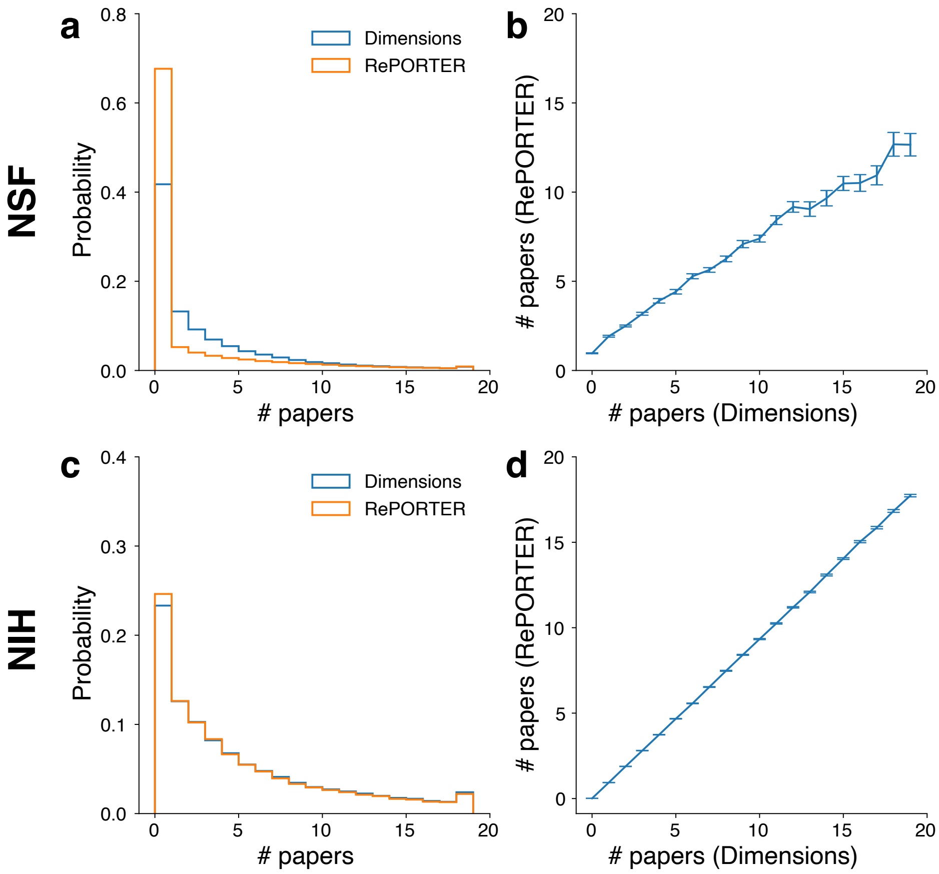

Figure 4.4. **Comparing papers supported by NSF and NIH grants based on RePORTER and Dimensions data.**

(a) The distribution of number of papers matched to$n = 169,086$ NSF grants. **(b)** Number of papers matched to each NSF grant reported by Dimensions and RePORTER data. Dimensions have more papers covered on average. Data are presented as mean values +/- SEM. **(c,d)** Same as (a,b) but for $n = 190,335$ NIH grants.

$$RCI_d^f = \frac{\text{\# papers in field } f \text{ consumed by domain d/ \# papers in } f}{\text{Total \# papers consumed by domain d / Total \# papers}}. \tag{4.1}$$

We find that the public uses of science are diverse, with many fields showing substantially specialized usage in public domains (Fig. 4.5d). Computer science, materials science, mathematics, and engineering (Fig. 4.5d, i-j) present substantially larger RCI values for patents than for government or news. By contrast, environmental science and geology (Fig. 4.5f,h) contribute relatively strongly in government and media documents compared to patents. Finally, physics, chemistry, medicine, and biology present a broader range of use (Fig. 4.5b-c, k-l). Among all fields, biology is the only one over-represented across all three channels, demonstrating a uniquely general relevance to these broad domains beyond science.

Social sciences, by contrast, exhibit a visibly different pattern of public use. The social sciences are strongly consumed in government and media domains while showing systematically low usage in patents (Fig. 4.5m-q). Economics sees especially strong government use, while psychology, sociology, and political science see relatively strong media use. Arts and humanities (philosophy, art and history, Fig. 4.5r-t) are relatively under-represented in all three domains.

Specialization in public use further appears at sub-domain levels (Fig. 4.6). For government, different agencies consume very different scientific research. For example, the U.S. Department of Treasury draws especially on economics and business research, the U.S. Department of Energy draws especially on geology and engineering, and the U.S. Department of Defense draws unusually on history. Different patenting fields further exhibit highly specialized relationships to specific scientific fields. By contrast, in media, while the Washington Post draws unusually heavily on political science research, mainstream

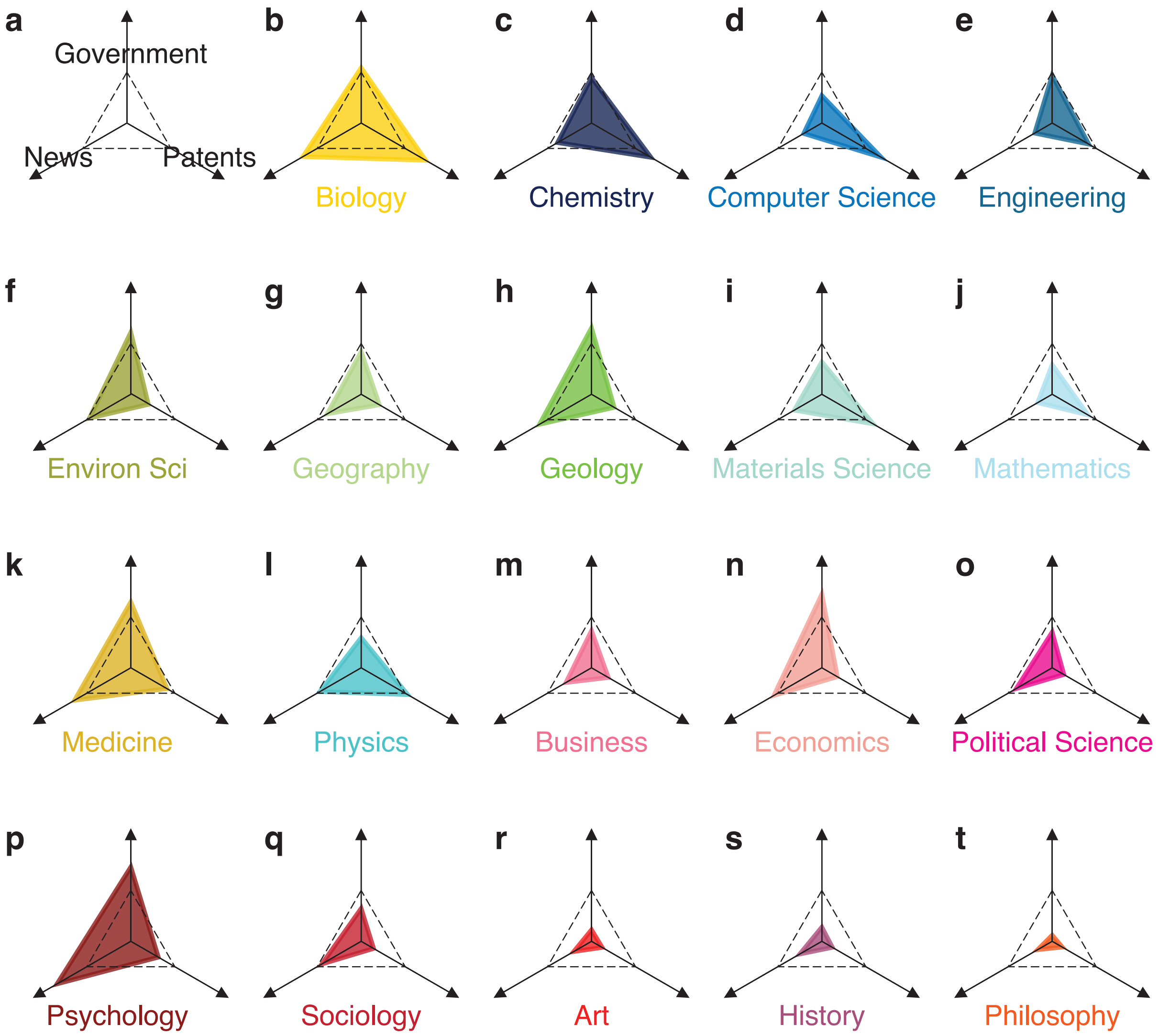

Figure 4.5. **Diversity in public use.**

Different scientific fields experience distinct and typically specialized public uses. **(a-t)** The usage metric RCI for the three public domains, presented for each field (b-t). The dashed triangles represent a null model where each paper has the same chance to be used (a). The color scheme highlights four high-level areas of research – the physical sciences, life sciences, social sciences, and ecology & earth sciences – following the four major clusters of science detected by [190] and suggesting commonalities in patterns of public use within these four areas.

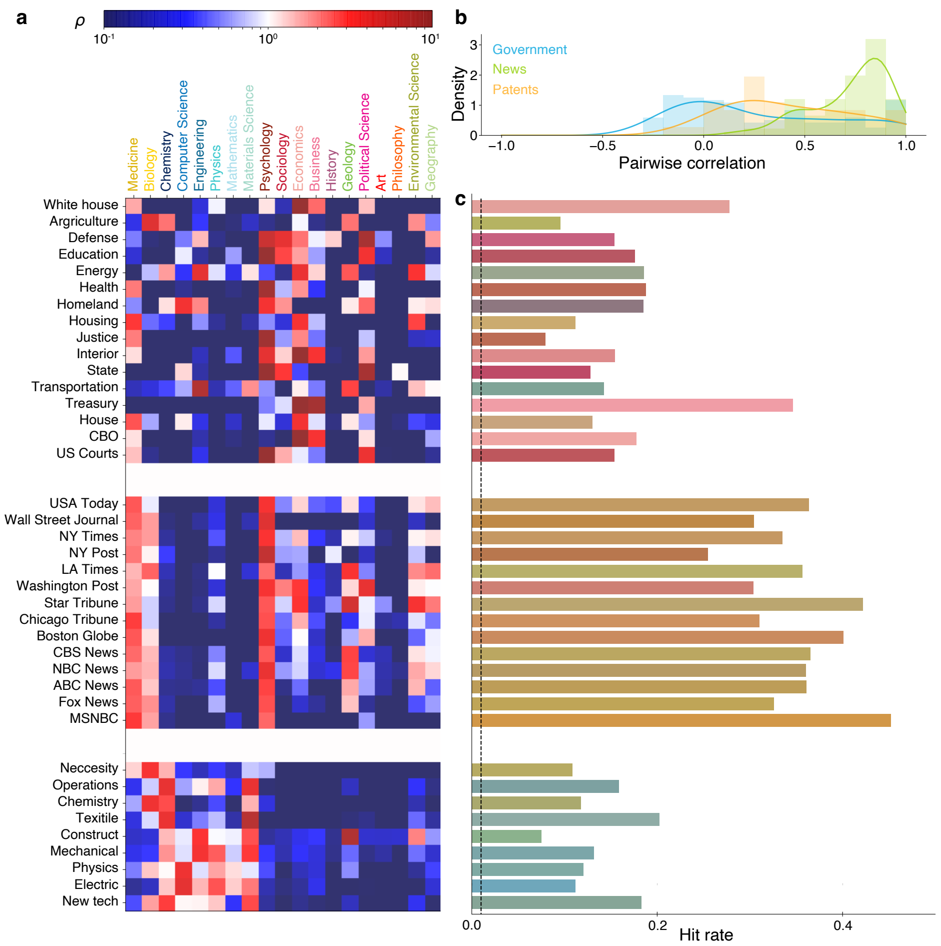

Figure 4.6. **Use of science across subdomains of government, news, and patents.**

(a) Patterns of consumption across subdomains. Subdomains are major departments and entities within the U.S. federal government, major news outlets for U.S. media, and top-level CPC (Cooperative patent classification) technology classes for patents. **(b)** The heterogeneity of scientific consumption across the subdomains in (a). **(c)** For every subdomain, paper hit rates are universally higher than the baseline (dashed line).

media sources in general are more consistent in the fields they report, with especially strong and widespread interest in medicine and psychology.

The specialization in public use is further accompanied by substantial differences in time lags in the use of science by the different public domains. Whereas the news media places a particular focus on very recent work, the government and inventive domains have wider reach into prior discovery. For example, in the news media, 63% of citations to scientific articles cover research papers published within the year. By contrast, government documents and patent inventions draw more widely over past work, with a median citation lag of 10 years between scientific publication and use (Fig. 4.7). We then examine the lags between papers and their use, both across fields and in the different public domains. Figure 4.7 presents these findings, with the time lag distribution for different fields (colored lines) presented in each panel together with the average across fields in that domain (black line). We see two types of heterogeneity here:

(1) We find substantial differences when comparing citation time lags across three public domains. Mainstream media covers mostly recent scientific research, with 63% of citations towards scientific papers published in the same year. At the same time, government documents and patent inventions show much longer lags between discovery and use, where a median citation time lag of 10 years, suggesting these two domains are more likely to draw on knowledge that has stood the test of time.

(2) Within each public domain, we also find field-level heterogeneities in the time lag. The median duration of policy citations to chemistry papers (15 years) is more than two times of that for economics publications (7 years) or for computer science publications (7 years).

To test the potential effects of heterogeneity in citation time lag on the field-level public use, we first expand the set of focal papers from the 10-year period (2005-2014) to a 30-year period (1985-2014). Importantly, while the public domains differ considerably in time lags, we find that the *RCI* comparisons are extremely similar when considering either

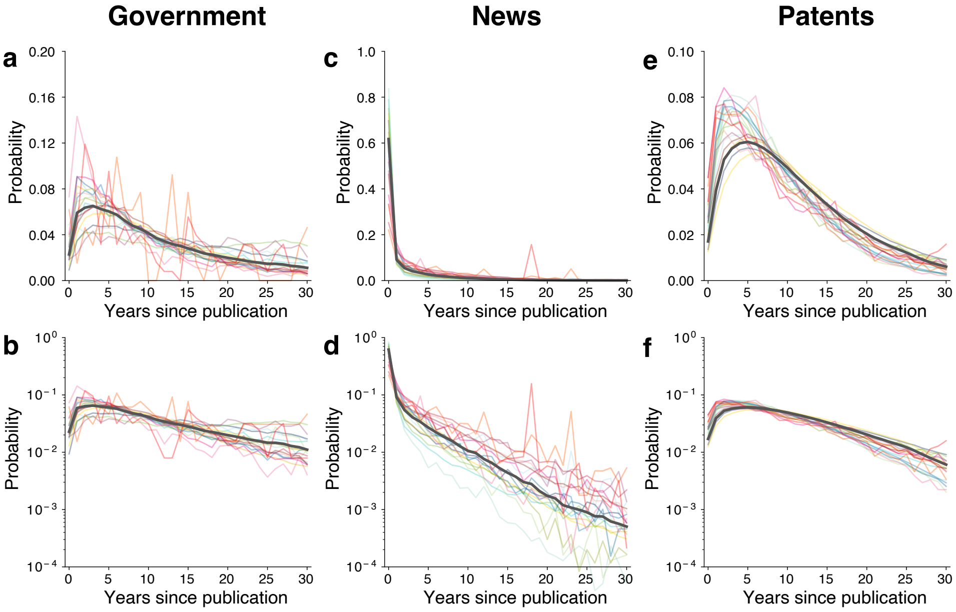

Figure 4.7. **The citation time lag between discovery and use.**

For visualization purposes, the top and bottom rows show results in linear-linear **(a,c,e)** and log-linear **(b,d,f)** scales respectively.

the recent decade of scientific publications (Fig. 4.5) or the stock of scientific publications over a substantially longer history (Fig. 4.8). Thus, although there is heterogeneity in lags, the *RCI* measure even over the 10-year period produces a similar picture as when looking over the much longer period.

Overall, these results highlight a large set of specialized relationships between specific domains of public use and specific fields of scientific research. From a public goods perspective, if we think of scientific fields as akin to a series of national parks, we see that each park is embedded in particular communities of public use. Collectively, these parks spread across diverse regions of knowledge and are accessed by diverse segments of

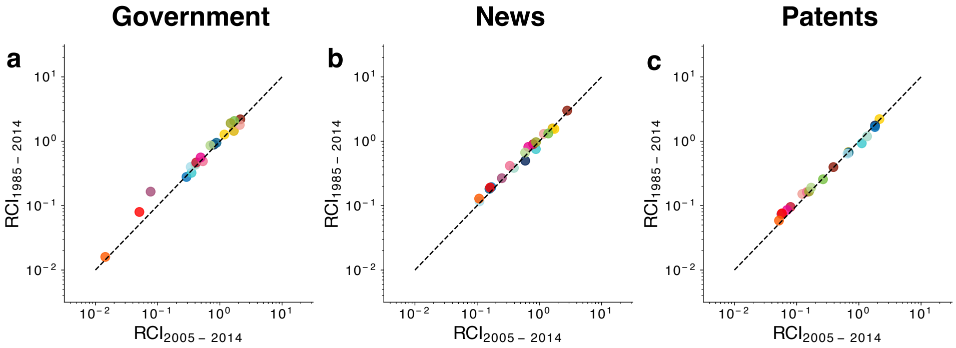

Figure 4.8. **Comparing RCI calculated on papers across different time periods (L0 level).**

the public. A few fields, and especially biology, receive visitors at relatively intense rates from a broad range of public domains – a "Yellowstone Park" of science.

4.4.2. Scientific impact and public use

Our second set of results examine whether the public domains tend to consume ideas that scientists themselves consider impactful. Longstanding arguments suggest that the public is not well equipped to evaluate science and may draw on poorly established scientific ideas, which would undermine the public good benefits of science [104, 152, 154]. Continuing the national parks metaphor, scientists may be primarily focused in a hard-to-reach backcountry, whereas the typical visitor may not have the tools to access this terrain nor gravitate to the same areas the scientists themselves consider attractive. To further examine public use, we therefore consider, at the article level, the alignment between public use and scientific use. While citations are widely used as a proxy for scientific impact [1, 2, 191–193], direct comparison of citation counts received by papers across time and field can be problematic without normalization [109, 131]. We therefore calculate citation

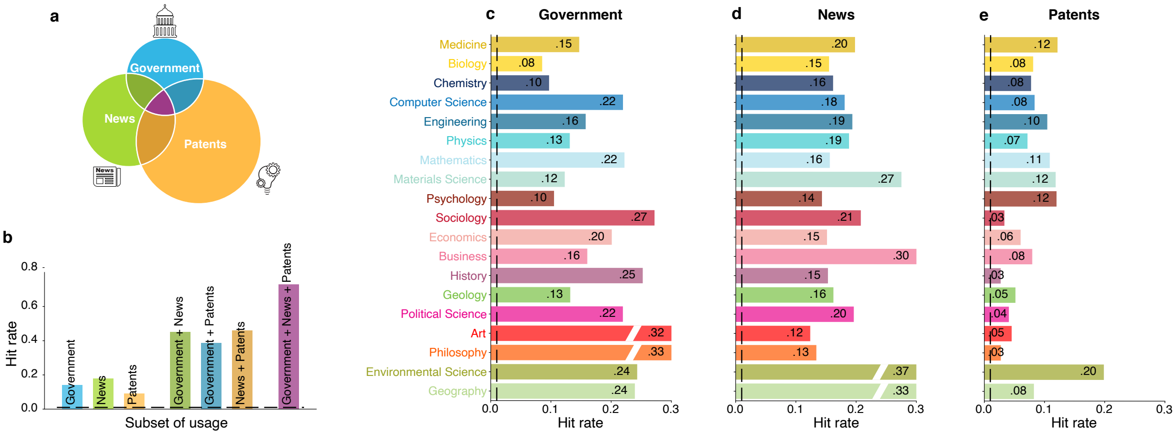

Figure 4.9. **Public use and scientific use.**

The public tends to consume exceptionally high impact science from all fields and in all three public domains, indicating alignment between public use and scientific use. **(a)** Usage by domain for papers published from 2005 to 2014. The area of each subset is proportional to the square root of the paper count in the corresponding public domain. **(b)** Hit rates for papers cited in at least one, two, or three public domains. Hit papers are defined as those receiving citation counts, within science, in the top 1% within the field and year. **(c-e)** Hit rates for each of the 19 fields consumed by government documents (c), news media (d) and patents (e). In all fields, and in all three domains, the consumed papers have hit rates within science many times larger than the baseline rate of 1% (dashed line).

percentiles for papers within the same publication year and field. Here following prior studies [15, 174, 194], we define 'hit papers' (also known as 'home runs') as papers ranking in the top 1% of citations received.

We find that papers referenced in public domains have a remarkably high likelihood of being hit papers within science (Fig. 4.9b). Papers cited by government documents, news or patents exhibit hit rates of 14.1%, 18.0% and 9.1%, respectively, all large multiples of the baseline rate of 1%. Further, papers referenced in the intersection of different domains tend to be exceptionally impactful in science. For papers referenced in two public domains, approximately half are hit papers. Papers referenced by both government documents and

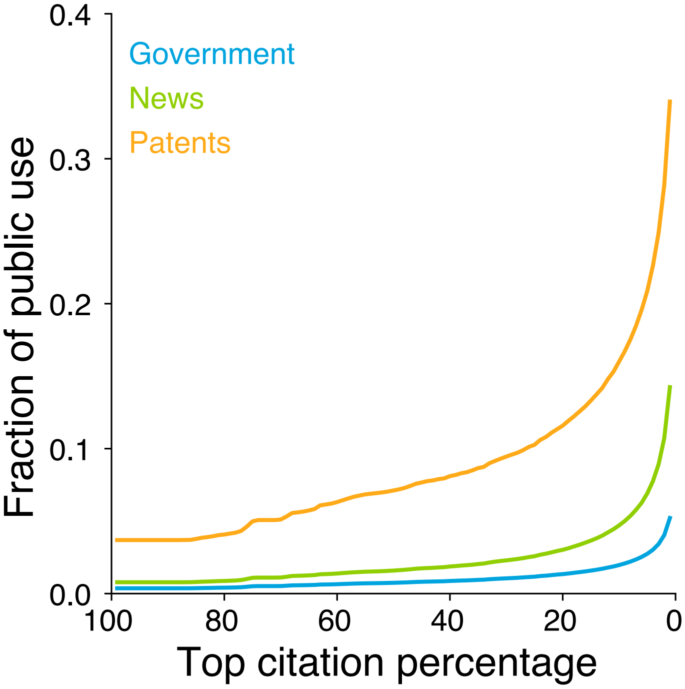

Figure 4.10. **The likelihood for papers ranked (top X% impact) to be cited in public domains.**

news media have a hit rate of 45.1%. The results are broadly similar if we examine the intersection between government documents and patents (38.7%) or news and patents (46.1%). A paper consumed in all three domains is a hit paper in science at a staggering 72.8 times the baseline rate. Reversing the exercise, we also see that, as the citation percentile of a paper rises, the probability for public use increases steeply, with extremely sharp increases at the very top of the citation distribution (Figs. 4.10, 4.11).

The use of high-impact papers is not only common across different public domains, it also appears universal across research areas. Papers covered by public domains tend to be highly cited in all scientific fields (Fig. 4.9c-e). These findings remain similar when varying the threshold for hit papers to the top 5% or 10% citations (Appendix). We

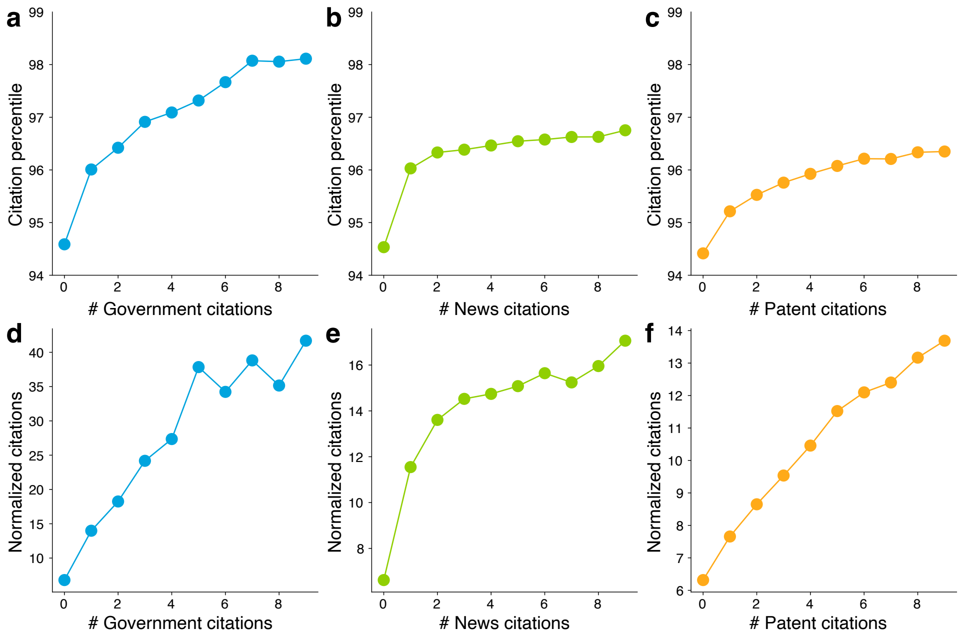

Figure 4.11. **Comparing impact in science and public within top 10% most cited papers.**

We use citation percentile **(a-c)** and normalized citations **(d-f)** within the same field and year as two measures of scientific impact.

further test robustness of these results by tuning the threshold from 1% to 5% or 10%. We also repeat our analyses for papers produced by U.S.-based researchers, arriving at the same conclusions. While government, media, and patenting documents may cite science for a variety of reasons and our reference-based measures are proxies for uses of science [103, 108, 125], we see that the science referenced in public domains is not in conflict with what scientists themselves consider important; rather, impactful papers defined by these communities show substantial overlap. This finding stands in contrast to concerns over knowledge gaps, where the government and media in particular may be poorly positioned to assess high impact scientific work or distinguish it from low impact scientific work [104, 152, 154, 163, 195]. Considering the findings, one may note that in each of these public domains, the initial step beyond science involves an intermediary – via the journalist in media, the inventor or other domain expert in patenting, the potential policy expert in government – all of whom may bring specialized capacities to bear in selecting what science they bring forth into their domain. The broader public use – among those who read a news article, use an invented product, or experience a policy – will then depend upon these intermediaries, who may help bridge the knowledge gap. Overall, the public use of science, while marked by substantial specialization in use across research areas, presents a striking universality, where diverse public domains all draw on the highest-impact scientific papers within each field.

4.4.3. Public use and public funding

We further fine-grain the 19 broad research fields of papers into 294 subfields as indexed by MAG, and calculate the *RCI* score for each subfield in a given public domain. We visualize each field's *RCI* values, locating each field within a common triangle to compare each field's tendency toward usage in specific public domains (Fig. 4.12a). Fields in social science as well as arts and humanities are mostly used in media and government, whereas fields in science and engineering spread out widely within the triangle, again highlighting the field-level specialization yet collective diversity in the public uses of science.

Together, these results raise a central question: To what degree does the funding input for science relate to the field's public use? The majority of scientific research is supported by public investment, which aims to advance not only science itself but also broader public interest [125]. The NSF, for example, formally introduced broader impacts

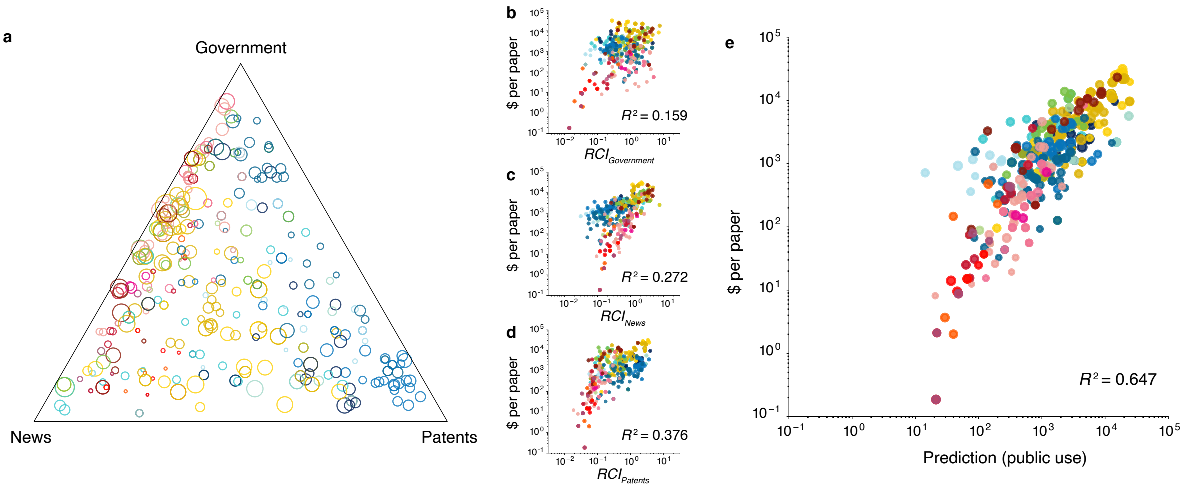

Figure 4.12. **Public use and public funding.**

Amidst enormous diversity in public use across fields and domains, scientific funding for a given field is closely aligned with the totality of its public use. **(a)** Ternary plot of RCI for 294 level-1 fields together, with the location of each field indicating its relative usage among the public domains. Circles are colored coded according to its parent field in Fig. 4.5, and circle sizes reflect overall usage. **(b-d)** Average funding per paper across fields is positively correlated with a field's RCI index in government (b), news (c) and patenting (d). The relationship remains significant when combined with control variables ($P < 0.001$ in OLS regressions controlling for the number of papers and parent field fixed effects, see Supplementary Table 2 for details). **(e)** Collectively, public uses beyond science strongly predict field level funding per paper.

as a key criterion for evaluating grant proposals in 1997. Here we focus on U.S.-funded projects and use D_5 to calculate the average funding per paper in a given subfield as a proxy for public investment costs per unit of output.

To understand the association between public use and funding for different scientific fields, we use linear regression models (ordinary least squares). We first note that all three RCI measures are highly skewed (Fig. 4.13a-c), prompting us we take the natural logarithm, $\ln RCI$, in our linear regressions (Fig. 4.13d-f). The same transformation is taken on the average funding per paper. The variables are defined as follows:

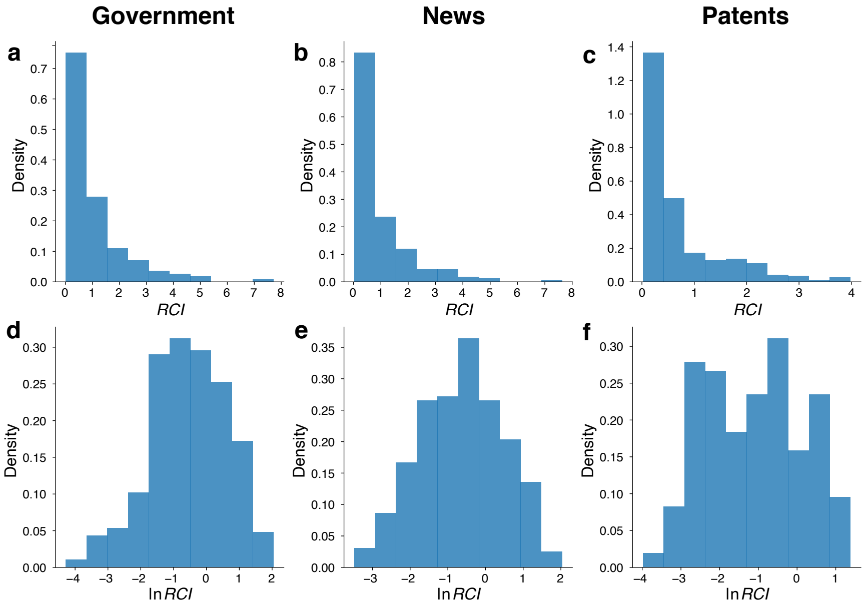

Figure 4.13. **Distribution of $\boldsymbol{RCI}$ index and transformations.**

(a-c) Histogram of RCI for level-1 fields in policy (a), news (b) and patent (c). All three distributions are highly skewed, prompting us to take appropriate transformations before regression analysis. **(a-c)** Histogram of $\ln RCI$ for level-1 fields in policy (d), news (e) and patent (f). The three distributions are closer to normal distributions after the transformation as compared with (a-c).

Dependent variable: $\ln\ Y_i$, defined as the natural logarithm of average funding per paper for the level-1 field i.

Predictors of interest: We examine the extent to which different impact measures can predict funding, including $\ln RCI_j$ for the three public domains, as well as $\ln c_i$, the natural logarithm of mean citations received for papers in that field. To include all data points in the regression, for the rare cases when an impact measure is 0, we add 1 to avoid 0s in the logarithm. We further include the natural logarithm of the number of papers published in the ten-year period, $\ln p_i$, as a control variable.

Fixed effects: To control for the possibility that fields under different broad categories may have specific funding and public use norms, we introduce F_{fi}, fixed effect terms for each level-0 field. Specifically, $F_{fi} = 1$ if the level-1 field i is a child field of the level-0 field f according to MAG's classification structure. Note that some level-1 fields belong to two level-0 fields simultaneously (e.g., mathematical physics is the child field of both mathematics and physics).

We start with bivariate regressions examining the relationship between each *RCI* (i.e., for government, media, or patenting) and average funding (Fig. 4.12b-d, Table 4.1, Models 1-3). That is,

$$\ln\ Y_i = \ \beta_j \ln RCI_{ji} + \varepsilon_i. \tag{4.2}$$

In multivariate regressions, we further include controls for heterogeneity in field size or parent field fixed effects (Table 4.2, Model 4-6).

We further investigate the joint predictive power of the three *RCI*s (Fig. 4.12e, Table 4.2, Model 7).

$$\ln\ Y_i = \sum_j \beta_j \ln RCI_{ji} + \varepsilon_i \tag{4.3}$$

which shows that each measure contributes independently and substantially to explaining the variation in funding.

Finally, we add further control variables into Models 8 (Table 4.2, Model 8).

$$\ln\ Y_i = \sum_j \beta_j \ln RCI_{ji} + \beta_p \ln p_i + \sum_f \beta_f F_{fi} + \varepsilon_i \tag{4.4}$$

We find that the public investment per paper differs dramatically across fields, spanning over five orders of magnitude. Yet comparing average funding per paper with *RCI* in each domain reveals substantial correlations between funding and the use of science across all three public domains, with $R^2 = 0.159$ for government, 0.272 for news, and 0.376 for patents (Fig. 3b-d, Methods, Supplementary Table 1). To further test if the uncovered correlation is due to the heterogeneity in field size or parent field, we add the number of papers in the subfield as well as parent field fixed effects (for the 19 higher-level fields) into the regression, finding the strong correlation with *RCI* persists ($P < 0.001$ in all three cases). Notably, across the three domains, the representation of subfields in government documents has the lowest predictive power for funding, suggesting that public investments in science better reflect the overall public interest captured by media or patents. We further include funding from non-governmental sources or focus on papers by US researchers only, finding our conclusions remain the same (Appendix).

Most strikingly, a simple linear regression model combining the three *RCI* values together yields a surprisingly high degree of agreement with funding, with an R^2 of 0.647 (Fig. 4.12, Table 4.2), providing at minimum a 72% increase in predictive power compared with using any of the three public domains alone. These results suggest that each public domain provides independent predictive power for understanding the allocation of public investment in science. The uncovered high predictive power of this analysis is especially striking given many complex factors and processes at work in appropriations, budget

[]

Model	(1)	(2)	(3)
VARIABLES			
Policy (RCI)	0.645***		
	(0.087)		
News (RCI)		0.880***	
		(0.084)	
Patent (RCI)			0.923***
			(0.070)
Observations	294	294	294
R2	0.159	0.272	0.376
F	55.17	108.8	175.9

Table 4.1. **Regression results for Models 1-3.**

Standard errors in parentheses. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

[]

Model	(4)	(5)	(6)	(7)	(8)
VARIABLES					
Policy (RCI)	0.502***			0.247***	0.208***
	(0.073)			(0.064)	(0.064)
News (RCI)		0.828***		0.728***	0.678***
		(0.076)		(0.067)	(0.074)
Patent (RCI)			0.659***	0.889***	0.591***
			(0.083)	(0.053)	(0.067)
# Paper (p)	0.243***	0.303***	0.160***		0.192***
	(0.052)	(0.048)	(0.053)		(0.043)
Level-0 fixed effect	Y	Y	Y		Y
Observations	294	294	294	294	294
R2	0.700	0.754	0.713	0.647	0.816
F	30.13	39.61	32.18	177.4	51.90

Table 4.2. **Regression results for Models 4-8.**

Standard errors in parentheses. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

setting, and grant review [3, 7, 52, 53, 55, 194]. Although each research field differs significantly in its relative role and contribution in science and beyond, the combination of their impacts beyond science powerfully predicts funding, suggesting that, ultimately, what the public uses, what scientists use, and what is funded are remarkably consistent.

4.4.4. The role of scientific funding, author characteristics, and journal prestige

We present extensive additional analysis to further inform the observed alignment between scientific use, public use, and funding. Before turning to these analyses, it is useful to emphasize that the paper's primary purpose (including the data assembly and analysis) is to test commonly held views that there is little or no alignment between the science and public spheres and that the methods are intended to draw together high-scale data and establish novel facts that speak to those hypotheses. The methodology is not intended to produce causal interpretation, which we consider an exciting avenue for future work deploying experimental or quasi-experimental methods. That said, adding in finer-grained analyses can reveal important, additional insights.

To further analyze the relationships between public use, funding, and impact, here we leverage variation at the paper level. This allows us to see whether the journal placement, author traits (like eminence), and paper-level funding predict public use, and whether such considerations can explain the large variation in public use across fields. For example, promotional advantages may increase public use – where greater financial resources for a research team, the prestige and marketing efforts of a journal, or perhaps individual eminence advantages may increase public attention to certain scientific papers.

First consider regressions at the paper level. The dependent variable is an indicator for whether a given paper is referenced in one of our public domains. The explanatory variables include an indicator for whether the paper is high impact (in the top 1% of citations in its field and publication year) and an indicator for whether the paper is supported by public funding. We then further include fixed effects for (a) the journal in which the paper appears (accounting for features like journal prestige as an information cue to the public, or variation in journals' capacity to market ideas); (b) individual author fixed effects, which can account for the author's capacity for self-promotion, eminence, or other personal factors; and (c) field fixed effects, which account for remaining differences across fields in their average public use once the journal, authors, impact, and funding of the specific paper have been taken into account. To capture author-level advantages, note that we deploy two alternatives: (i) the average h-index of all authors affiliated with a paper and (ii) author-level fixed effects. Due to the large sample size of our data (116.6M author-paper pairs in 2005-2014), we run (ii) by randomly sampling 1% of the authors in our data for computational efficiency.

Supplementary Table 8 presents the regression results. First, we see a very large effect of being a high impact paper. That is, even conditional on whether the paper is publicly-funded, in a given field, in a given journal, by a particular authors, etc., public use is sharply predicted by impact within science. This indicates that the link between impact within science and public use is highly robust, and not simply a matter of journal prestige, author prestige, or funding status.

Second, we also see evidence that is consistent with some promotional advantages. Public funding does suggest greater public use at the paper level, conditional on the other controls. That is, even conditional on a paper being high impact in a given field (and in a given journal and by a particular author, etc.) public funding has additional positive predictive capacity for public use, although this paper-level result is quite weak compared to being a high-impact paper.

The journal dimension can be further seen in Fig. 4.14. Here we separately consider groups of journals within each field. Specifically, we calculate a measure of expected hit rate at the journal level, defined as the frequency for papers on a particular journal to be a "hit" paper (i.e., top 1% cited within the same year and field). The journal-level expected hit rate therefore offers an approximation for journal impact. To test if the observed pattern between public use and impact exists outside the very top journals, we construct a list of the top 10 journals in each L0 field (ranked by the journal-level expected hit rate), which includes journals such as *Nature*, *Science*, and *PNAS*. We then repeat our results in Fig. 4.9b looking outside the top 10 journals in each field. We still see a strong linkage between the paper hit rate within science and public use.

We also find that fields continue to vary massively in how often their papers are taken up, and in a similar way regardless of paper-level considerations. One way to see this is in Fig. 4.15, where we separated papers by their funding status and looked at their *RCI* measured separately. For each L1 field, we calculate the *RCI* measures separately based on funded papers, unfunded papers, and unfunded hit papers. These *RCI* measures prove highly correlated with each other. Therefore, fields with relatively high public attention (high RCI measures) are high attention regardless of whether the specific paper was publicly funded. This suggests that public funding at the paper level is not really what's driving the relative attention to the field. Yet another way to see this is to look at the relationship between the raw *RCI* measure and the field fixed effects in the above regression, as shown in Fig. 4.16. Here we see that the field differences appear broadly similar both in the raw data and when net of all the controls (journal, individual fixed effects, paper funding, and paper impact).

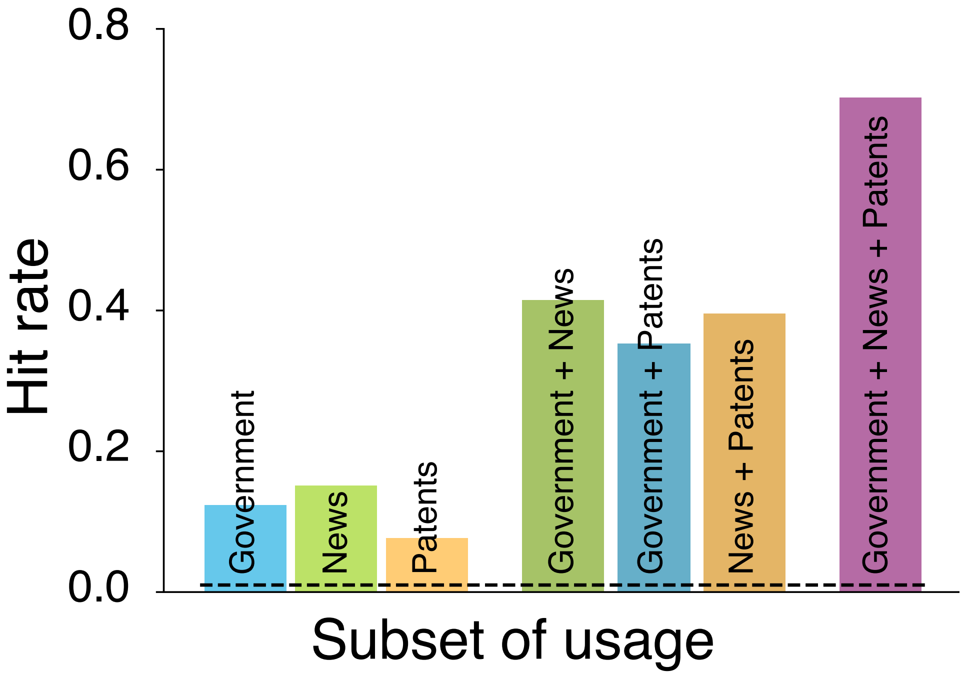

Figure 4.14. **Hit rates for papers cited in at least one, two, or three public domains.**

This figure repeats the analysis of Fig. 4.9b but dropping the top 10 journals by impact in each field.

Altogether, while a paper's funding may advance attention to it, as may a high-prestige journal or individual, the field-level variation in public attention appears robust to these considerations. Ultimately, this paper-level analysis suggests that the large variation in attention to different research areas appears to be primarily a feature of the area itself, rather than the specific promotion opportunities from a journal, scientist, or funding.

4.5. Concluding remarks

One source of this alignment could be that science follows the public interest. For example, scientists may prioritize or innately share areas of interest, such as COVID-19,

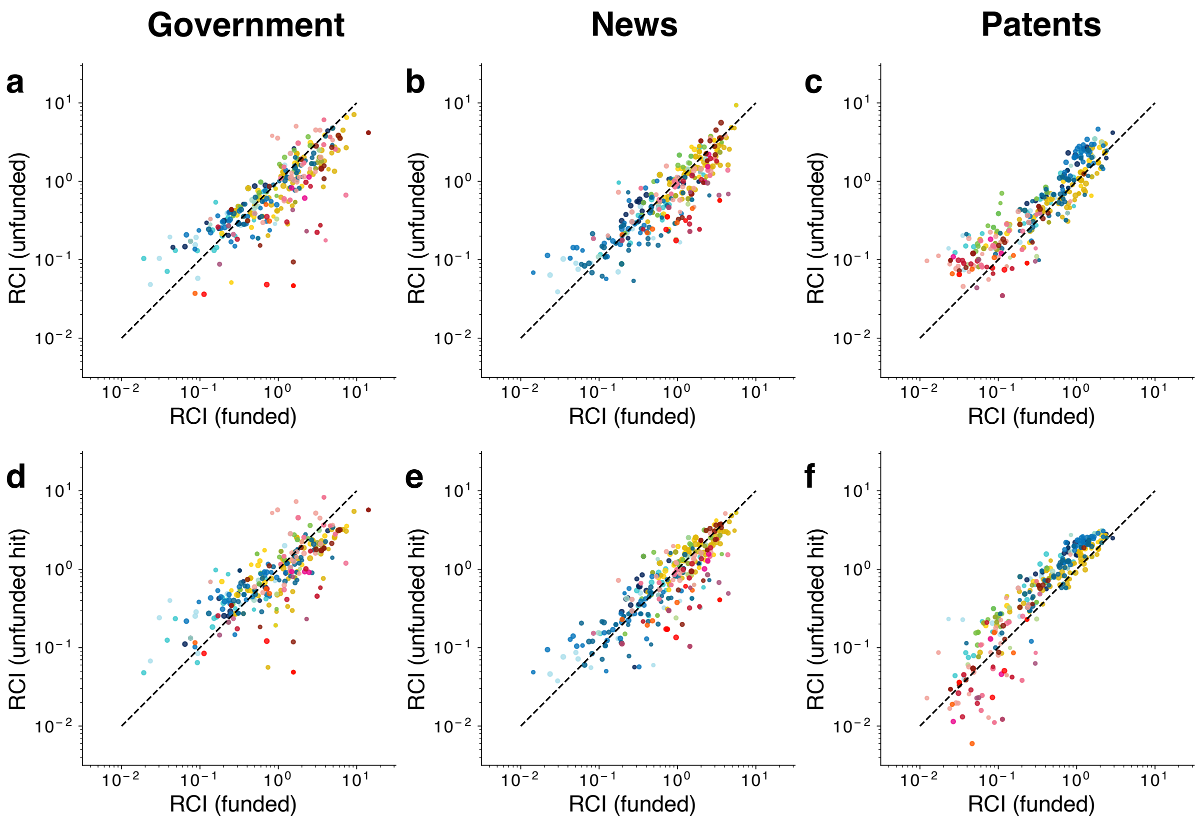

Figure 4.15. **Comparing RCI values between funded papers with unfunded papers (a-c) and unfunded hit papers (d-f) at the L1 level.**

This figure compares relative attention to different fields when looking at funded papers in that field versus unfunded papers in that field. The relative attention to a field tends to be similar either way, suggesting that public attention to a field is not being driven by paper-level funding.

where there is enormous public demand for solutions and where scientific attention has surged [177, 196, 197]. Another source could be that some scientists or science institutions are especially good at promoting their interests to the public, influencing what the public sees and funds. For example, one may wonder if high-prestige journals, eminent authors, or funding for a paper drive attention to specific research. To test this, we further consider fine-grained, paper-level regressions that include journal fixed effects, author fixed effects,

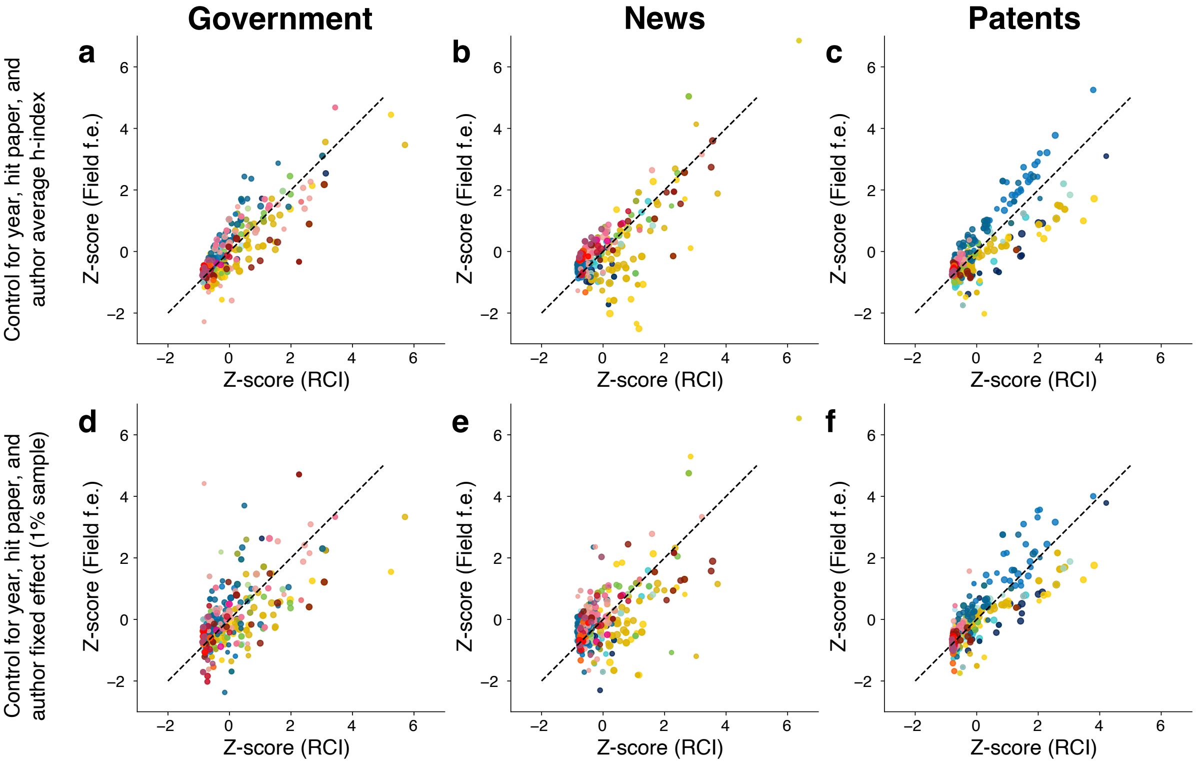

Figure 4.16. **Comparing RCI values and field fixed effects at the L1 level.**

We examine average attention to a field through field fixed effects in a regression. The regression seeks to explain public attention to a paper based on whether a paper was funded, the paper's citation impact, author characteristics, journal fixed effect, and field effects. Author-level advantages are controlled for by author average h-index at the paper level **(a-c)** and author fixed effects on 1% random sample of all authors **(d-f)**.

and paper-level funding indicators. We find that the results are very similar, regardless of these controls (Fig. 4.14, 4.15, 4.16, Table 4.3). Indeed, the relative attention to different fields (Fig. 4.5), the alignment between public use and high-impact science (Fig. 4.9), and the alignment with public funding (Fig. 4.12), all appear robust after accounting for journal placement, the scientists who produced the work, or the funding status of the specific paper. Thus, while some scientists, journals, or funders may have advantages

[]

Model	(1)	(2)	(3)	(4)	(5)	(6)
	Government uses		News uses		Patent uses	
	(dummy)		(dummy)		(dummy)	
VARIABLES						
Hit paper (indicator)	0.045***	0.036***	0.107***	0.135***	0.233***	0.258***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.002)
Funded (indicator)	0.002***	0.001***	0.009***	0.011***	0.026***	0.033***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Author h-index	0.000***		0.000***		0.000***	
	(0.000)		(0.000)		(0.000)	
Author fixed effect		Y		Y		Y
Level-1 field fixed effect	Y	Y	Y	Y	Y	Y
Year fixed effect	Y	Y	Y	Y	Y	Y
Journal fixed effect	Y	Y	Y	Y	Y	Y
R2	0.036	0.190	0.079	0.250	0.121	0.265
F	702.0	13.41	2216.8	70.18	3949.6	91.03

Table 4.3. **Regression results for public uses of science.**

Standard errors in parentheses. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

in reaching the public, the forms of alignment we see appear primarily as features of a research area, rather than the specific promotion opportunities from a journal, scientist, or funding. More generally, numerous mechanisms, institutional factors, and policies may be at work in producing, increasing, or reducing use and alignment, and unpacking these mechanisms is an exciting area for future work.

Altogether, the analysis probes quantitatively key features of the public use and funding of science. Measuring the usage of scientific research outside science itself, we uncover enormous diversity and specialization in how different fields of scientific inquiry are linked to different public domains. Yet, despite these differences, the different public domains (and subdomains) universally draw on highly cited papers within science, indicating that public use is strongly aligned with what scientists themselves consider impactful. And, critically, the public usage of scientific fields across the diverse domains provides simple yet powerful predictors for the level of public investment in each field.

Note that, although the three domains each represent an important dimension of the public space, they do not cover all domains that science may impact. Even within each of the three domains we studied, there may be consumption of science through channels that go beyond our datasets. For example, scientists and their ideas can appear through television, in congressional testimony, and in private sector consulting. Scientific ideas may also enter industry and government through social networks, through the hiring of scientists, and through influencing managerial practices (Fig. 4.3), which may augment and alter perspectives on the public use of specific research fields. While there is much still to explore, this paper introduces a quantitative framework to examine public uses of science at the individual paper level, both across all scientific fields and diverse public domains, revealing individually specialized and collectively diverse uses, universality in impact, and a remarkable alignment between the funding of science and its public use.

As society's support of science depends on a public goods model [144, 146], and as legislators have called for more transparency in the usage and value of scientific funding [198], the framework developed in this paper provides an empirical tool, offering quantitative evidence to inform discussions around public interest features of science. The allocation of science funding involves chains of decisions by individuals and groups with different perspectives and priorities. These considerations range from legislative committees and the goals of individual political representatives, to funding agency leaders, to within-agency mechanisms that often incorporate insights from scientists, interacting in a complex process that must bridge across distinct communities. As such, one might expect a substantial disconnect between what is eventually funded and forms of public interest – metaphorically, funding of public parks in ways weakly related to public use. Yet, despite the massive diversity in the public uses of science and a complex funding process, there is remarkable alignment in the end result. What the public uses and what scientists themselves use are closely consistent. And the funding of science closely tracks quantifiable public use. These results suggest the connections between the ivory tower and the real world appear more aligned than is commonly imagined.

CHAPTER 5

Conclusion

The rapid growth of scientific research and its increasing complexity have created unprecedented opportunities and challenges for understanding and managing the scientific enterprise. Understanding the key mechanisms underlying the full spectrum of scientific achievements and impact, from success to failure, from scientific influence to broad impact, carries growing importance for identifying and nurturing new ideas, talents, and paradigms across a wide range of domains.

This dissertation illustrates the power of interdisciplinary tools to unearth how different individual, social, and environmental factors can promote (or inhibit) scientific and technological progress. Chapter 2 would not have been possible without the modeling framework rooted in complex systems, which is further combined with large-scale data analytics. By mimicking how future attempts build on those past, this simple yet powerful mechanistic model speaks directly to a wide range of literature, including innovations, human dynamics, and learning theories, capable of covering the key predictions by existing models in its limiting cases. More importantly, following the idea of phase transitions – a fundamental concept from statistical physics – the model separates failure dynamics failure into regions of stagnation or progression, making four different empirically testable predictions. At the same time, the empirical validity and practical relevance are only possible through the three large-scale datasets in science, security, and startups. These corpora are among the largest in their respective settings, yet at the same time, also differ dramatically in their scope, scale, definition, and temporal resolution, allowing us to systematically interrogate the theoretical framework we developed. Chapter 3 builds directly on canonical science of science tools such as citation networks analysis, which is further applied to high-resolution tracking data of scientific and policy landscapes during the pandemic. These results not only offer a novel quantitative framework to quantify the interconnectedness of the COVID-19 policy responses, but also present some of the most urgently needed empirical evidence to inform a large set of global issues and debates by the time of writing, ranging from the role of science and journals in this crisis to the US President Trump's decision to defund WHO amidst the global crisis. Further, the frameworks developed are not limited to COVID-19, but can be generalized to study real-time policy responses to various emergent threats, from natural disasters to human conflicts to more, holding omnipresent relevance to the scientific community and beyond. Chapter 4 represents an emerging opportunity for the science of science research in the Big Data era. By leveraging the latest computational tools in information retrieval, we are now able to systematically link scientific publications and researchers to a wide range of complementary features that are not tracked in traditional citation databases, from upstreaming public funding to downstream public uses in government policy, media news, and marketplace applications. Such linkages offer an unprecedented opportunity to examine the role of science in the broader public space, which not only relates to the long-standing interest in the sociology and economics of science literature, but also serves as the foundational idea of the institutional architectures to support science through public investment in many countries. Together, these three examples illustrate the power of integrating fundamental ideas and tools from physical, computational, and social sciences.

From a broader computational social science perspective, science and innovation serve as a powerful lens to examine broader social processes. As shown in this dissertation, a deeper understanding of science as a model system can produce highly generalizable methods and conclusions to complement and enrich data-driven insights across a diverse set of complex social systems. For example, by developing a simple yet general model of repeated failures, Chapter 2 offers a universal quantitative basis not only for scientists obtaining NIH grants but also for business innovators achieving startup exits and terrorist organizations launching fatal attacks. Given the ubiquity of failure and knowledge spillover in various social systems, the techniques and insights presented here may prove useful for our understanding of other social systems, ranging from artistic and cultural productions to public policy and media attention to market competition and human conflict.

The endeavors made in this work are certainly not complete. Indeed, as simple as many of the measurement frameworks we established here, they have achieved remarkable success in uncovering hidden patterns within and beyond the knowledge ecosystems, suggesting a high degree of regularity and predictability underlying the seemingly noisy and random data. One interesting extension is to build and analyze a more comprehensive modeling framework of learning for benchmarking and diagnosing the fundamental constraints in successful productivity. The extant literature on the essential tension between exploration and exploitation, from areas as diverse as sociology, psychology, economics, physics, and innovation, highlights that neither pure exploration nor pure exploitation is ideal, implying the existence of an optimal progress trajectory when exploration and exploitation are combined. A deeper understanding of this optimal trajectory, both theoretically and empirically, carries direct implications for the diagnosis, improvement, and planning of all innovative activities where learning dynamics play a role. It would also be interesting to examine the collective dynamics behind scientific and technological frontiers, building on canonical insights across diverse domains such as organizational search, cultural evolution, crowd wisdom, and technology management. For example, are there reproducible patterns governing the evolution of technological frontiers? Are there signals for an impending scientific breakthrough? What kinds of teams or strategies are most effective in shaping the rate and direction of technological progress and disruptive innovations? Finally, given the substantial time gap between scientific production and public uses, one can imagine a data-driven framework to quantify and forecast how today's scientific discoveries enable tomorrow's technological and commercial inventions. Ultimately, an improved ability to understand and predict the use and applications of new scientific knowledge would improve and accelerate scientific and technological progress, the fundamental engine of economic growth and human prosperity.

References

[1] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[2] Dashun Wang and Albert-László Barabási. *The science of science.* Cambridge University Press, 2021.

[3] Yian Yin, Yang Wang, James A Evans, and Dashun Wang. Quantifying the dynamics of failure across science, startups and security. *Nature*, 575(7781):190–194, 2019.

[4] Danielle Li and Leila Agha. Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, 348(6233):434–438, 2015.

[5] Paula E Stephan. *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA, 2012.

[6] Cary P Gross, Gerard F Anderson, and Neil R Powe. The relation between funding by the national institutes of health and the burden of disease. *New England Journal of Medicine*, 340(24):1881–1887, 1999.

[7] Donna K Ginther, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, and Raynard Kington. Race, ethnicity, and nih research awards. *Science*, 333(6045):1015–1019, 2011.

[8] Yongwook Paik. Serial entrepreneurs and venture survival: Evidence from us venture-capital-financed semiconductor firms. *Strategic Entrepreneurship Journal*, 8(3):254–268, 2014.

[9] Paul Gompers, Anna Kovner, Josh Lerner, and David Scharfstein. Performance persistence in entrepreneurship. *Journal of Financial Economics*, 96(1):18–32, 2010.

[10] JP Eggers and Lin Song. Dealing with failure: Serial entrepreneurs and the costs of changing industries between ventures. *Academy of Management Journal*, 58(6):1785–1803, 2015.

[11] Steven N Kaplan and Josh Lerner. Venture capital data: Opportunities and challenges. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*. University of Chicago Press, 2016.

[12] Grace S Walsh, James A Cunningham, et al. Business failure and entrepreneurship: emergence, evolution and future research. *Foundations and Trends® in Entrepreneurship*, 12(3):163–285, 2016.

[13] National Consortium for the Study of Terrorism and Responses to Terrorism (START). *Global Terrorism Database [Data file]*. 2018.

[14] Tim Harford. *Adapt: Why success always starts with failure*. Farrar, Straus and Giroux, 2011.

[15] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[16] Benjamin F Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

[17] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[18] Lu Liu, Yang Wang, Roberta Sinatra, C Lee Giles, Chaoming Song, and Dashun Wang. Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559(7714):396, 2018.

[19] Yanqing Hu, Shlomo Havlin, and Hernán A Makse. Conditions for viral influence spreading through multiplex correlated social networks. *Physical Review X*, 4(2):021031, 2014.

[20] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[21] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[22] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

[23] R Dean Malmgren, Daniel B Stouffer, Andriana SLO Campanharo, and Luis A Nunes Amaral. On universality in human correspondence activity. *Science*, 325(5948):1696–1700, 2009.

[24] Linda Argote. *Organizational learning: Creating, retaining and transferring knowledge*. Springer Science & Business Media, 2012.

[25] Sim B Sitkin. Learning through failure: the strategy of small losses. *Research in organizational behavior*, 14:231–266, 1992.

[26] Louis E Yelle. The learning curve: Historical review and comprehensive survey. *Decision sciences*, 10(2):302–328, 1979.

[27] John M Dutton and Annie Thomas. Treating progress functions as a managerial opportunity. *Academy of management review*, 9(2):235–247, 1984.

[28] George P Huber. Organizational learning: The contributing processes and the literatures. *Organization science*, 2(1):88–115, 1991.

[29] Mark D Cannon and Amy C Edmondson. Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. *Long range planning*, 38(3):299–319, 2005.

[30] Brian A Jacob and Lars Lefgren. The impact of research grant funding on scientific productivity. *Journal of public economics*, 95(9-10):1168–1177, 2011.

[31] Danielle Li, Pierre Azoulay, and Bhaven N Sampat. The applied value of public investments in biomedical research. *Science*, 356(6333):78–81, 2017.

[32] James McNerney, J Doyne Farmer, Sidney Redner, and Jessika E Trancik. Role of design complexity in technology improvement. *Proceedings of the National Academy of Sciences*, 108(22):9008–9013, 2011.

[33] Jerker Denrell and James G March. Adaptation as information restriction: The hot stove effect. *Organization Science*, 12(5):523–538, 2001.

[34] Linda Argote, Sara L Beckman, and Dennis Epple. The persistence and transfer of learning in industrial settings. *Management science*, 36(2):140–154, 1990.

[35] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

[36] Robert K Merton. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470–486, 1961.

[37] Neil Johnson, Spencer Carran, Joel Botner, Kyle Fontaine, Nathan Laxague, Philip Nuetzel, Jessica Turnley, and Brian Tivnan. Pattern in escalations in insurgent and terrorist activity. *Science*, 333(6038):81–84, 2011.

[38] Pablo Martin de Holan and Nelson Phillips. Remembrance of things past? the dynamics of organizational forgetting. *Management science*, 50(11):1603–1613, 2004.

[39] Allen Newell and Paul S Rosenbloom. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, 1(1981):1–55, 1981.

[40] John R Anderson. Acquisition of cognitive skill. *Psychological review*, 89(4):369, 1982.

[41] John F Muth. Search theory and the manufacturing progress function. *Management Science*, 32(8):948–962, 1986.

[42] Theodore P Wright. Factors affecting the cost of airplanes. *Journal of the aeronautical sciences*, 3(4):122–128, 1936.

[43] Richard F Bass. *Stochastic processes*, volume 33. Cambridge University Press, 2011.

[44] V Zaburdaev, S Denisov, and J Klafter. Lévy walks. *Reviews of Modern Physics*, 87(2):483, 2015.

[45] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[46] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[47] Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

[48] Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

[49] Nassim Nicholas Taleb. *The black swan: The impact of the highly improbable*, volume 2. Random house, 2007.

[50] Harold Stanley Heaps. *Information retrieval, computational and theoretical aspects.* Academic Press, 1978.

[51] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[52] Kevin J Boudreau, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, 62(10):2765–2783, 2016.

[53] Lindell Bromham, Russell Dinnage, and Xia Hua. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684, 2016.

[54] Albert Banal-Estanol, Inés Macho-Stadler, and David Pérez Castrillo. Key success drivers in public research grants: Funding the seeds of radical innovation in academia? 2016.

[55] Athen Ma, Raúl J Mondragón, and Vito Latora. Anatomy of funded research in science. *Proceedings of the National Academy of Sciences*, 112(48):14760–14765, 2015.

[56] James G March. Exploration and exploitation in organizational learning. *Organization science*, 2(1):71–87, 1991.

[57] Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

[58] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[59] Daniel M Wegner. Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior*, pages 185–208. Springer, 1987.

[60] Diane Wei Liang, Richard Moreland, and Linda Argote. Group versus individual training and group performance: The mediating role of transactive memory. *Personality and social psychology bulletin*, 21(4):384–393, 1995.

[61] Linda Argote, Bill McEvily, and Ray Reagans. Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management science*, 49(4):571–582, 2003.

[62] Robert K Merton et al. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[63] Alexander M Petersen, Woo-Sung Jung, Jae-Suk Yang, and H Eugene Stanley. Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23, 2011.

[64] Pierre Azoulay, Toby Stuart, and Yanbo Wang. Matthew: Effect or fable? *Management Science*, 60(1):92–109, 2013.

[65] Jian Huang, Seyda Ertekin, and C Lee Giles. Efficient name disambiguation for large-scale databases. In *European conference on principles of data mining and knowledge discovery*, pages 536–544. Springer, 2006.

[66] Helen Shen. Inequality quantified: Mind the gender gap. *Nature News*, 495(7439):22, 2013.

[67] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature News*, 504(7479):211, 2013.

[68] Tiantian Yang and Howard E Aldrich. Who's the boss? explaining gender inequality in entrepreneurial teams. *American Sociological Review*, 79(2):303–327, 2014.

[69] Linda Argote, Chester A Insko, Nancy Yovetich, and Anna A Romero. Group learning curves: The effects of turnover and task complexity on group performance. *Journal of Applied Social Psychology*, 25(6):512–529, 1995.

[70] Charles D Bailey. Forgetting and the learning curve: A laboratory study. *Management science*, 35(3):340–352, 1989.

[71] Rita Gunther McGrath. Falling forward: Real options reasoning and entrepreneurial failure. *Academy of Management review*, 24(1):13–30, 1999.

[72] Amy C Edmondson. Strategies for learning from failure. *Harvard business review*, 89(4):48–55, 2011.

[73] Dean A Shepherd. Learning from business failure: Propositions of grief recovery for the self-employed. *Academy of management Review*, 28(2):318–328, 2003.

[74] Jerker Denrell. Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, 14(3):227–243, 2003.

[75] Kristina B Dahlin, You-Ta Chuang, and Thomas J Roulet. Opportunity, motivation, and ability to learn from failures and errors: Review, synthesis, and ways to move forward. *Academy of Management Annals*, 12(1):252–277, 2018.

[76] Ji-Yub Kim and Anne S Miner. Vicarious learning from the failures and near-failures of others: Evidence from the us commercial banking industry. *Academy of Management Journal*, 50(3):687–714, 2007.

[77] Amy C Edmondson. Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *The Journal of Applied Behavioral Science*, 40(1):66–90, 2004.

[78] Peter M Madsen. These lives will not be lost in vain: Organizational learning from disaster in us coal mining. *Organization Science*, 20(5):861–875, 2009.

[79] Joel AC Baum and Kristina B Dahlin. Aspiration performance and railroads' patterns of learning from train wrecks and crashes. *Organization Science*, 18(3):368–385, 2007.

[80] Pamela R Haunschild and Bilian Ni Sullivan. Learning from complexity: Effects of prior accidents and incidents on airlines' learning. *Administrative science quarterly*, 47(4):609–643, 2002.

[81] Peter M Madsen and Vinit Desai. Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. *Academy of Management Journal*, 53(3):451–476, 2010.

[82] Ferdinand K Levy. Adaptation in the production process. *Management Science*, 11(6):B–136, 1965.

[83] Barbara Levitt and James G March. Organizational learning. *Annual review of sociology*, 14(1):319–338, 1988.

[84] Linda Argote and Dennis Epple. Learning curves in manufacturing. *Science*, 247(4945):920, 1990.

[85] Devendra Sahal. A theory of progress functions. *AIIE Transactions*, 11(1):23–29, 1979.

[86] Peter Roberts. A theory of the learning process. *Journal of the Operational Research Society*, 34(1):71–79, 1983.

[87] Jeff Shrager, Tad Hogg, and Bernardo A Huberman. A graph-dynamic model of the power law of practice and the problem-solving fan-effect. *Science*, 242:414–416, 1988.

[88] Aaron Clauset and Kristian Skrede Gleditsch. The developmental dynamics of terrorist organizations. *PloS one*, 7(11):e48633, 2012.

[89] Avraham N Kluger and Angelo DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254, 1996.

[90] Harold Asher. *Cost-quantity relationships in the airframe industry.* PhD thesis, The Ohio State University, 1956.

[91] ERFW Crossman. A theory of the acqusition of speed-skill. *Ergonomics*, 2(2):153–166, 1959.

[92] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, 128(1):11–45, 1987.

[93] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[94] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H Strogatz. The dynamics of correlated novelties. *Scientific reports*, 4:5890, 2014.

[95] Iacopo Iacopini, Staša Milojević, and Vito Latora. Network dynamics of innovation processes. *Physical review letters*, 120(4):048301, 2018.

[96] Daniel A Levinthal. Adaptation on rugged landscapes. *Management science*, 43(7):934–950, 1997.

[97] George S Snoddy. Learning and stability: a psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, 10(1):1, 1926.

[98] John Laird, Paul Rosenbloom, and Allen Newell. *Universal subgoaling and chunking: The automatic generation and learning of goal hierarchies*, volume 11. Springer Science & Business Media, 2012.

[99] Frank E Ritter and Lael J Schooler. The learning curve. *International encyclopedia of the social and behavioral sciences*, 13:8602–8605, 2001.

[100] Vittorio Loreto, Vito DP Servedio, Steven H Strogatz, and Francesca Tria. Dynamics on expanding spaces: modeling the emergence of novelties. In *Creativity and universality in language*, pages 59–83. Springer, 2016.

[101] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[102] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007.

[103] National Research Council. *Using science as evidence in public policy*. National Academies Press, 2012.

[104] Charles Percy Snow. *Science and Government*. Harvard University Press, Cambridge, MA, 2013.

[105] R. Haunschild and L. Bornmann. How many scientific papers are mentioned in policy-related documents? an empirical investigation using web of science and altmetric data. *Scientometrics*, 110(3):1209–1216, 2017.

[106] S. S. Jasanoff. Contested boundaries in policy-relevant science. *Social Studies of Science*, 17(2):195–230, 1987.

[107] S. Hilgartner. The dominant view of popularization: Conceptual problems, political uses. *Social Studies of Science*, 20(3):519–539, 1990.

[108] C. H. Weiss. The many meanings of research utilization. *Public Administration Review*, 39(5):426–431, 1979.

[109] Nathan Caplan. The two-communities theory and knowledge utilization. *American behavioral scientist*, 22(3):459–470, 1979.

[110] M. Zastrow. Open science takes on the coronavirus pandemic. *Nature*, 581(7806):109–110, 2020.

[111] Alex John London and Jonathan Kimmelman. Against pandemic research exceptionalism. *Science*, 368(6490):476–477, 2020.

[112] Alfred HJ Kim, Jeffrey A Sparks, Jean W Liew, Michael S Putman, Francis Berenbaum, Alí Duarte-García, Elizabeth R Graef, Peter Korsten, Sebastian E Sattui, and Emily Sirotich. A rush to judgment? rapid reporting and dissemination of results and its consequences regarding the use of hydroxychloroquine for covid-19. *Annals of Internal Medicine*, 172:819–821, 2020.

[113] http://help.overton.io/en/.

[114] https://blog.overton.io/blog/2020/04/23/who-are-the-covid-19-guidelines-policy-citing/.

[115] E. Adie. Personal communication. 2020.

[116] C. Herzog, D. Hook, and S. Konkiel. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395, 2020.

[117] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infectious Diseases*, 20:533–534, 2020.

[118] B. Kale, H. V. Siravuri, H. Alhoori, and M. E. Papka. Predicting research that will be cited in policy documents. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 389–390. Association for Computing Machinery.

[119] Andy Tattersall and Christopher Carroll. What can altmetric.com tell us about policy citations of research? an analysis of altmetric.com data for research articles from the university of sheffield. *Frontiers in Research Metrics and Analytics*, 2:9, 2018.

[120] H. Yu, X. Cao, T. Xiao, and Z. Yang. How accurate are policy document mentions? a first look at the role of altmetrics database. *Scientometrics*, 125:1517–1540, 2020.

[121] R. Newson, L. Rychetnik, L. King, A. Milat, and A. Bauman. Does citation matter? research citation in policy documents as an indicator of research impact–an australian obesity policy case-study. *Health Research Policy and Systems*, 16(1):55, 2018.

[122] S. Vilkins and W. J. Grant. Types of evidence cited in australian government publications. *Scientometrics*, 113(3):1681–1695, 2017.

[123] L. Bornmann, R. Haunschild, and W. Marx. Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents? *Scientometrics*, 109(3):1477–1495, 2016.

[124] C. D. Willis, B. Riley, L. Stockton, S. Viehbeck, S. Wutzke, and J. Frank. Evaluating the impact of applied prevention research centres: Results from a modified delphi approach. *Research Evaluation*, 26(2):78–90, 2017.

[125] L. Bornmann. What is societal impact of research and how can it be assessed? a literature survey. *Journal of the American Society for Information Science and Technology*, 64(2):217–233, 2013.

[126] I. Tahamtan and L. Bornmann. Altmetrics and societal impact measurements: Match or mismatch? a literature review. *El profesional de la información (EPI)*, 29(1):e290102, 2020.

[127] Vladimír Šucha and Marta Sienkiewicz. *Science for Policy Handbook*. Elsevier, 2020.

[128] D. Contandriopoulos, M. Lemire, J. L. Denis, and E. Tremblay. Knowledge exchange processes in organizations and policy arenas: A narrative systematic review of the literature. *The Milbank Quarterly*, 88(4):444–483, 2010.

[129] L. Bornmann. Field classification of publications in dimensions: A first case study testing its reliability and validity. *Scientometrics*, 117(1):637–640, 2018.

[130] C. Herzog and B. K. Lunn. Response to the letter 'field classification of publications in dimensions: A first case study testing its reliability and validity'. *Scientometrics*, 117(1):641–645, 2018.

[131] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the*

National Academy of Sciences, 105(45):17268–17272, 2008.

[132] H Holden Thorp. Why who? *Science*, 368(6489):341, 2020.

[133] Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. Variation in government responses to covid-19. *Blavatnik School of Government Working Paper*, pages BSG–WP–2020/032, 2020.

[134] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L.Y. Huang, A. Hultgren, E. Krasovich, and P. Lau. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584(7820):262–267, 2020.

[135] J. Tollefson. How trump damaged science-and why it could take decades to recover. *Nature*, 586(7828):190–194, 2020.

[136] Benjamin Disraeli. *Inaugural address delivered to the University of Glasgow Nov. 19, 1873.* Longmans, Green, and Co., London, 1873.

[137] Robert K Merton. *The sociology of science: Theoretical and empirical investigations.* University of Chicago press, 1973.

[138] Michael Gibbons. *The new production of knowledge: The dynamics of science and research in contemporary societies.* Sage, 1994.

[139] Joel Mokyr. *The gifts of Athena: Historical origins of the knowledge economy.* Princeton University Press, 2002.

[140] Henry Etzkowitz and Loet Leydesdorff. The dynamics of innovation: from national systems and “mode 2” to a triple helix of university–industry–government relations. *Research policy*, 29(2):109–123, 2000.

[141] Committee on Prospering in the Global Economy of the 21st Century, National Academy of Sciences, and National Academy of Engineering Institute of

Medicine. *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* National Academies Press, 2014.

[142] Jonas Hjort, Diana Moreira, Gautam Rao, and Juan Francisco Santini. How research affects policy: Experimental evidence from 2,150 brazilian municipalities. Report 0898-2937, National Bureau of Economic Research, 2019.

[143] Thomas Jefferson. No patent on ideas: letter to isaac mcpherson. *August*, 13:1813, 1813.

[144] Kenneth Arrow. Economic welfare and the allocation of resources for invention. the rate and direction of inventive activity: economic and social factors. *N. Bureau*, 1962.

[145] Joseph E Stiglitz. Knowledge as a global public good. *Global public goods*, 1(9):308–326, 1999.

[146] Paula E Stephan. The economics of science. *Journal of Economic literature*, 34(3):1199–1235, 1996.

[147] John Jewkes. *The sources of invention.* Springer, 1969.

[148] Michael Gibbons and Ron Johnston. The roles of science in technological innovation. *Research policy*, 3(3):220–242, 1974.

[149] Ralph Landau, Nathan Rosenberg, and National Academy of Engineering. *The positive sum strategy: Harnessing technology for economic growth.* National Academies Press, 1986.

[150] Edwin Mansfield. Academic research and industrial innovation. *Research policy*, 20(1):1–12, 1991.

[151] Alvin K Klevorick, Richard C Levin, Richard R Nelson, and Sidney G Winter. On the sources and significance of interindustry differences in technological opportunities. *Research policy*, 24(2):185–205, 1995.

[152] William N Dunn. The two-communities metaphor and models of knowledge use: An exploratory case survey. *Knowledge*, 1(4):515–536, 1980.

[153] Laurence E Lynn. Knowledge and policy: The uncertain connection. 1978.

[154] Réjean Landry, Moktar Lamari, and Nabil Amara. The extent and determinants of the utilization of university research in government agencies. *Public Administration Review*, 63(2):192–205, 2003.

[155] John Langrish, Michael Gibbons, William G Evans, and Frederic Raphael Jevons. *Wealth from knowledge: Studies of innovation in industry*. Springer, 1972.

[156] Elaine Hatfield. Proxmire's golden fleece award. *Relationship Research News (Newsletter of the International Association for Relationship Research)*, 4:5–9, 2006.

[157] Tom Coburn. *The National Science Foundation: Under the microscope*. Senator Tom Coburn, 2011.

[158] Matt Ridley. *The evolution of everything: How new ideas emerge*. HarperCollins, 2015.

[159] Terence Kealey. The case against public science. *Cato Unbound*, 5, 2013.

[160] Ronald G Havelock. *Planning for innovation through dissemination and utilization of knowledge*. Center for Research on Utilization of Scientific Knowledge, Institute for . . . , 1979.

[161] R. Landry, N. Amara, and M. Lamari. Climbing the ladder of research utilization: Evidence from social science research. *Science Communication*, 22(4):396–422, 2001.

[162] S. Jasanoff. *States of knowledge: the co-production of science and the social order*. Routledge, 2004.

[163] Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5), 2013.

[164] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. Do "altmetrics" correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019, 2015.

[165] José Luis Ortega. Blogs and news sources coverage in altmetrics data providers: a comparative analysis by country, language, and subject. *Scientometrics*, 122(1):555–572, 2020.

[166] Stefanie Haustein. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, 108(1):413–423, 2016.

[167] Lutz Bornmann. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of informetrics*, 8(4):895–903, 2014.

[168] Lutz Bornmann. Validity of altmetrics data for measuring societal impact: A study using data from altmetric and f1000prime. *Journal of informetrics*, 8(4):935–950, 2014.

[169] Cassidy R Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9):2037–2062, 2017.

[170] Ruben Gaetani and Matteo Li Bergolis. The economic effects of scientific shocks. *Unpublished Manuscript*, 2015.

[171] Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 2020.

[172] Diana Hicks, Tony Breitzman, Dominic Olivastro, and Kimberly Hamilton. The changing composition of innovative activity in the us—a portrait based on patent analysis. *Research policy*, 30(4):681–703, 2001.

[173] Martin Meyer. Does science push technology? patents citing scientific literature. *Research policy*, 29(3):409–434, 2000.

[174] Mohammad Ahmadpoor and Benjamin F Jones. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587, 2017.

[175] Lee Fleming, Hillary Greene, G Li, Matt Marx, and Dennis Yao. Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141, 2019.

[176] Benjamin F Jones and Lawrence H Summers. A calculation of the social returns to innovation. Report, National Bureau of Economic Research, 2020.

[177] Yian Yin, Jian Gao, Benjamin F Jones, and Dashun Wang. Coevolution of policy and science during the pandemic. *Science*, 371(6525):128–130, 2021.

[178] Kuansan Wang, Zhihong Shen, Chi-Yuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45, 2019.

[179] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020.

[180] https://www.govinfo.gov.

[181] https://github.com/allenai/science-parse.

[182] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[183] Euan Adie and William Roe. Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1):11–17, 2013.

[184] Matt Marx and Aaron Fuegi. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. Report, National Bureau of Economic Research, Inc, 2020.

[185] http://www.patentsview.org.

[186] Daniel W Hook, Simon J Porter, and Christian Herzog. Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23, 2018.

[187] https://www.aaas.org/programs/r-d-budget-and-policy/federal-rd-budget-dashboard.

[188] https://report.nih.gov.

[189] https://federalreporter.nih.gov.

[190] Martin Rosvall and Carl T Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[191] Eugene Garfield and Robert K Merton. *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. Wiley New York, 1979.

[192] Derek J De Solla Price. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515, 1965.

[193] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[194] Yang Wang, Benjamin F Jones, and Dashun Wang. Early-career setback and future career impact. *Nature communications*, 10(1):1–10, 2019.

[195] Senthil Selvaraj, Durga S Borkar, and Vinay Prasad. Media coverage of medical journals: do the best articles make the news? *PLoS One*, 9(1), 2014.

[196] Ryan Hill, Yian Yin, Carolyn Stein, Dashun Wang, and Benjamin F Jones. Adaptability and the pivot penalty in science. *Available at SSRN 3886142*, 2021.

[197] Holly Else. How a torrent of covid science changed research publishing–in seven charts. *Nature*, 588(7839):553–554, 2020.

[198] Hearing: The science of science and innovation policy. *Committee on Science and Technology*, 2010.

APPENDIX A

Robustness checks

A.1. Robustness checks for Chapter 2

A.1.1. Definition of success and failure

We vary our definition of success and failure across different datasets. For D_1 we remove all renewal/resubmission successes and only focus on new applications, finding our conclusions are not affected by resubmissions (Fig. A.1).

For D_2 we vary the definition of success for a startup. Previously we have considered IPO and high-value M&A as success. Similar with hit papers defined in science of science, we define high-value M&As as those with transaction value ranking top 1% among all transactions in the same year. We vary this definition to top 5% transactions or exclude all M&As (Fig. A.2), finding our conclusions still hold. One problem with our definition for success is that it does not include ventures that could already be considered successful despite not having had an IPO or being acquired. To this end, we collected a list of unicorn companies, defined as privately held startup companies valued at over 1 billion, from CB Insights website, yielding 121 companies in our sample, which can be linked through company names. Overall we find such cases are relatively rare. We also test our conclusions by removing these cases from the unsuccessful group, or re-defining them as successful attempts. In both cases we find our results remain the same (Fig. A.2).

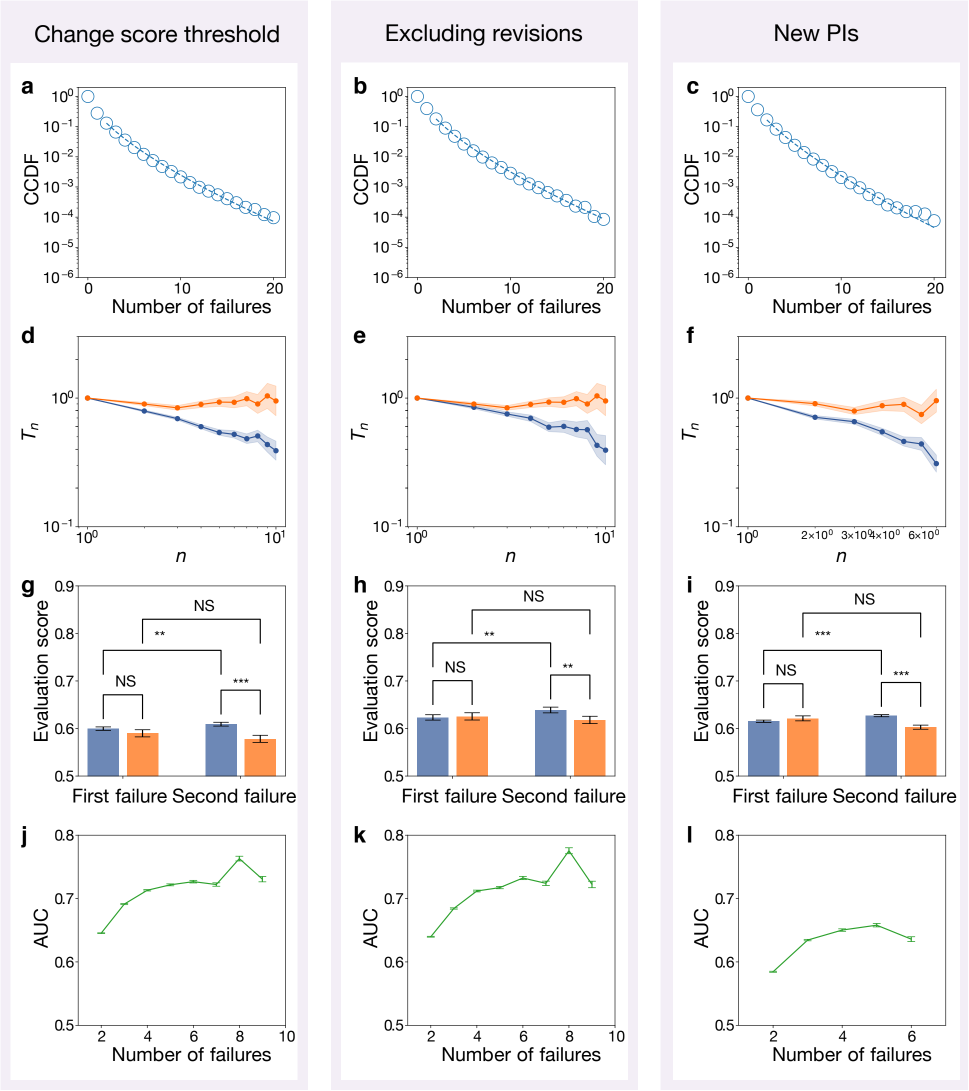

Figure A.1. **Robustness check on D_1.**

(Caption next page.)

For D_3 we tried variants by expanding unsuccessful groups to all samples or restricting successful groups to human-target samples only (Fig. A.3). Both variants yield similar results. We also vary the threshold in our data, changing our definition of successful group

(Previous page.) **a-c**, Failure streak as we change score threshold to 55 (**a**), exclude revisions as successes (**b**) and only focus on new PIs without previous R01 grants (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns as we change score threshold to 55 (**d**), exclude revisions as successes (**e**) and only focus on new PIs without previous R01 grants (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics as we change score threshold to 55 (**g**, $n = 768,\ 189,\ 686,\ 170$ respectively), exclude revisions as successes (**h**, $n = 252,\ 145,\ 216,\ 123$ respectively) and only focus on new PIs without previous R01 grants (**i**, $n = 1164,\ 308,\ 1530,\ 334$ respectively). The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for g,h and 3 for i) appear indistinguishable in first failures (two-sided t-test, $P = 0.242,\ 0.819,\ 0.289$) but quickly diverge in second failures (two-sided t-test, $P = 3.40 \times 10^{-4},\ 3.40 \times 10^{-2},\ 9.70 \times 10^{-7}$). The success group also shows significant performance improvement (one-sided t-test, $P = 4.23 \times 10^{-2},\ 3.04 \times 10^{-2},\ 1.92 \times 10^{-4}$), which is absent for the non-success group (one-sided t-test, $P = 0.863,\ 0.754,\ 0.997$). The centers and error bars denote the mean and s.e.m. **j-l**, AUC score of predicting ultimate success as we change score threshold to 55 (**j**), exclude revisions as successes (**k**) and only focus on new PIs without previous R01 grants (**l**). The centers and error bars of AUROC scores denote the mean and s.e.m calculated from 10-fold cross validation over 50 randomized iterations. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$.

as organizations that killed at least 5, 10 and 100 people in a terrorist attack (Fig. A.3). We find the patterns hold the same.

A.1.2. Threshold for being inactive in the system

The definition of unsuccessful group depends on the threshold for inactive in the system. In main text we set up the threshold as 5 years, i.e. if one does not appear in the system for the last 5 years, we consider such cases as drop-out samples. To test the effect of this threshold, here we repeat our main results for 3 years and 7 years (Fig. A.4), respectively. We find all our results are robust as we tune this criterion.

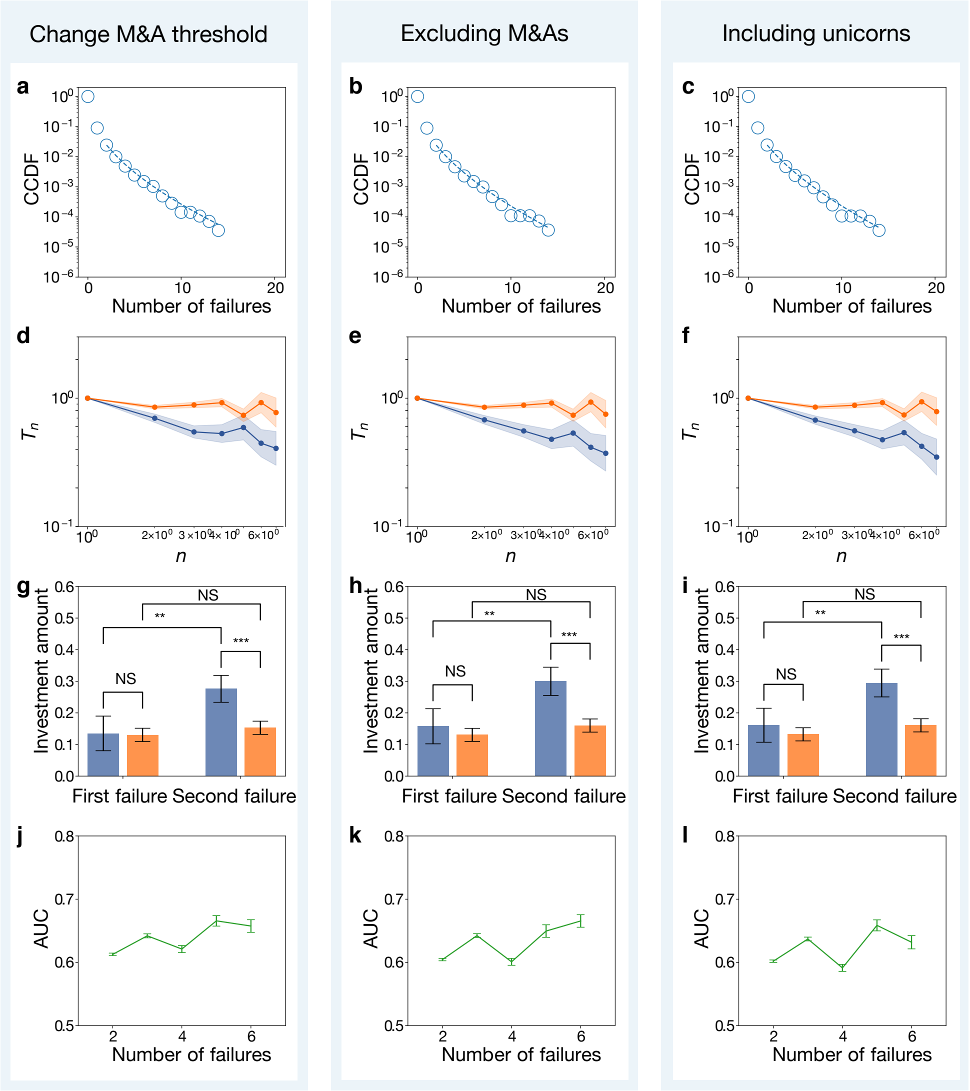

Figure A.2. **Robustness check on D_2.**

(Caption next page.)

A.1.3. Effect of overall success rate

It is also important to keep in mind that the success rate may go up and down over time. Here we control for the overall success rates across our three datasets and test its potential impact on our results. More specifically, we renormalize our empirical data by

(Previous page.) **a-c**, Failure streak as we change threshold of high-value M&A to 5% (**a**), exclude M&As as successes (**b**) and classify unicorns as successes (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns as we change threshold of high-value M&A to 5% (**d**), exclude M&As as successes (**e**) and include unicorns as successes (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics as we change threshold of high-value M&A to 5% (**g**, $n = 251,\ 1304,\ 243,\ 1284$ respectively), exclude M&As as successes (**h**, $n = 248,\ 1335,\ 237,\ 1315$ respectively) and include unicorns as successes (**i**, $n = 257,\ 1330,\ 244,\ 1311$ respectively). The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 3) appear indistinguishable in first failures (two-sided t-test, $P = 0.937,\ 0.647,\ 0.620$) but quickly diverge in second failures (two-sided t-test, $P = 9.92 \times 10^{-3},\ 4.94 \times 10^{-3},\ 6.33 \times 10^{-3}$). The success group also shows significant performance improvement (one-sided t-test, $P = 2.16 \times 10^{-2},\ 2.37 \times 10^{-2},\ 2.77 \times 10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.224,\ 0.158,\ 0.167$). The centers and error bars denote the mean and s.e.m. **j-l**, AUC score of predicting ultimate success as we change threshold of high-value M&A to 5% (**j**), exclude M&As as successes (**k**) and include unicorns as successes (**l**). The centers and error bars of AUROC scores denote the mean and s.e.m calculated from 10-fold cross validation over 50 randomized iterations. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$.

weighing different samples by success rate to ensure that each year has effectively the same success rate. For example, for samples from the successful group ending in year y, we count the total number of successes and failures in that year, defined as $S(y)$ and $F(y)$. We then calculate the weight of each sample as $w \equiv (F + S)/S$, i.e., the inverse of the overall success rate. This is equivalent to resampling within all successful cases, with the sampling probability proportional to the inverse of the success rate. To this end, the weighted sum of each year's success should be $S/w = (F + S)$, or proportional to the total number of samples in the same year.

We then repeat all of our main measurements using the renormalized samples. As shown in Fig. A.5, all of the main predictions made in our paper hold the same. This suggests that even though intelligence agencies may improve their ongoing detection of

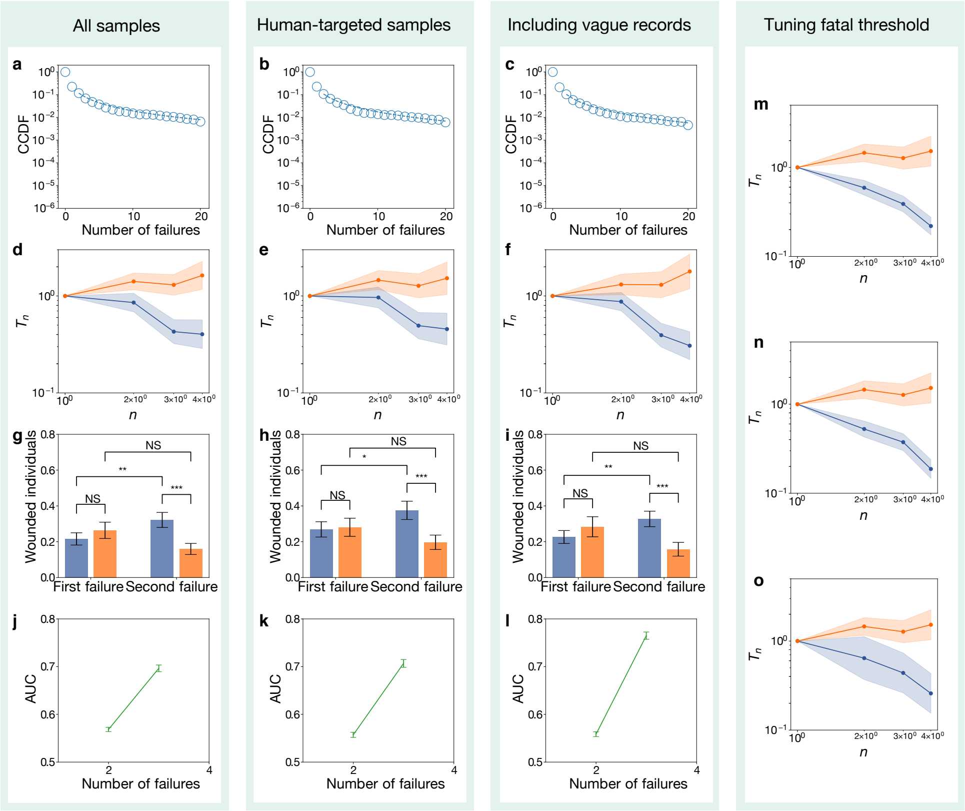

Figure A.3. **Robustness check on D_3.**

(Caption next page.)

terror attacks, congress may decrease (or increase) its annual budget for science, and economic cycles may increase or reduce the companies with successful exits, these changes are smooth in time, and do not affect the conclusions.

(Previous page.) **a-c**, Failure streak as we focus on all samples (**a**), samples of human-targeted attacks (**b**) and include vague data on fatality (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns as we focus on all samples (**d**), samples of human-targeted attacks (**e**) and include vague data on fatality (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics as we focus on all samples (**g**, $n = 231,\ 231,\ 229,\ 232$ respectively), samples of human-targeted attacks (**h**, $n = 176,\ 173,\ 173,\ 174$ respectively) and include vague data on fatality (**i**, $n = 227,\ 147,\ 225,\ 148$ respectively). The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 2) appear indistinguishable in first failures (two-sided t-test, $P = 0.400,\ 0.859,\ 0.395$) but quickly diverge in second failures (two-sided t-test, $P = 2.08 \times 10^{-3},\ 6.70 \times 10^{-3},\ 3.76 \times 10^{-3}$). The success group also shows significant performance improvement (one-sided t-test, $P = 2.55 \times 10^{-2},\ 5.65 \times 10^{-2},\ 3.77 \times 10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.970,\ 0.901,\ 0.967$). The centers and error bars denote the mean and s.e.m. **j-l**, AUC score of predicting ultimate success as we focus on all samples (**j**), samples of human-targeted attacks (**k**) and include vague data on fatality (**l**). The centers and error bars of AUROC scores denote the mean and s.e.m calculated from 10-fold cross validation over 50 randomized iterations. **m-o**, Temporal scaling patterns as we change the threshold for the success group as fatal attacks that killed at least 5 (**m**), 10 (**n**) and 100 (**o**) people. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$.

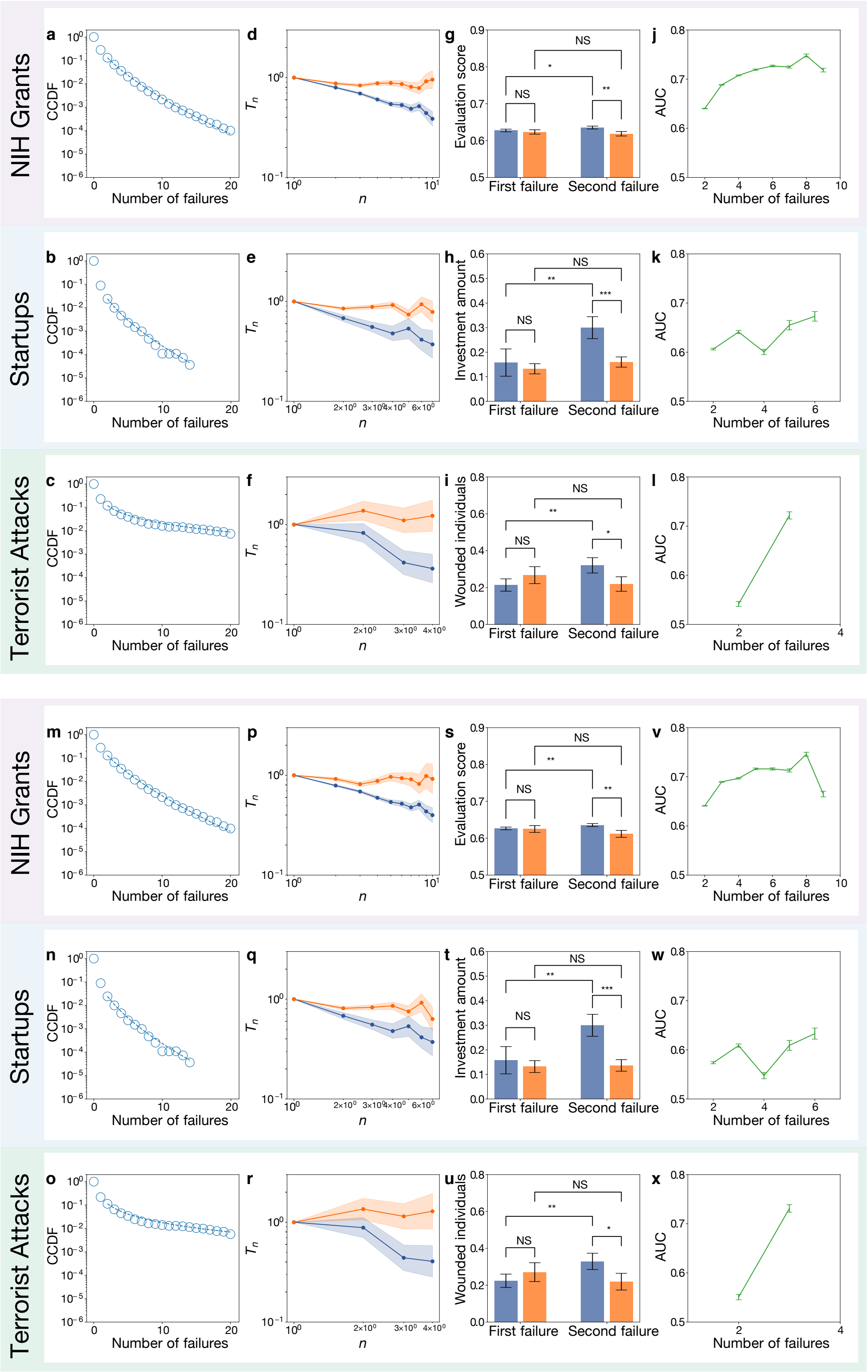

Figure A.4. **Robustness check on definition of non-success group.**

(Caption next page.)

(Previous page.) **a-l**, Robustness check as we change the threshold of inactivity to 3 years. **a-c**, Failure streak in science (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics in science (**g**, $n = 641,\ 231,\ 578,\ 190$ respectively), entrepreneurship (**h**, $n = 248,\ 1332,\ 237,\ 1312$ respectively) and security (**i**), $n = 238,\ 198,\ 236,\ 199$ respectively. The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable in first failures (two-sided t-test, $P = 0.566,\ 0.671,\ 0.349$) but quickly diverge in second failures (two-sided t-test, $P = 2.09 \times 10^{-2},\ 4.95 \times 10^{-3},\ 7.77 \times 10^{-2}$). The success group also shows significant performance improvement (one-sided t-test, $P = 7.03 \times 10^{-2},\ 2.37 \times 10^{-2},\ 2.32 \times 10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.717,\ 0.176,\ 0.786$). The centers and error bars denote the mean and s.e.m. **j-l**, AUC score of predicting ultimate success in science (**j**), entrepreneurship (**k**) and security (**l**). The centers and error bars of AUROC scores denote the mean and s.e.m calculated from 10-fold cross validation over 50 randomized iterations. **m-x**, As in **a-l** but using 7 years as the threshold of inactivity. Sample sizes are **s**: $n = 620,\ 101,\ 559,\ 76$; **t**: $n = 248,\ 977,\ 237,\ 989$; **v**: $n = 216,\ 152,\ 214,\ 153$ respectively. P-values in **u-w** are $P = 0.883,\ 0.671,\ 0.456$; $P = 2.25 \times 10^{-2},\ 1.38 \times 10^{-3},\ 8.34 \times 10^{-2}$; $P = 4.59 \times 10^{-2},\ 2.37 \times 10^{-2},\ 3.33 \times 10^{-2}$; $P = 0.838,\ 0.446,\ 0.775$. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$.

(Previous page.) **a-l**, Robustness check as we change the threshold of inactivity to 3 years. **a-c**, Failure streak in science (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics in science (**g**, $n = 641,\ 231,\ 578,\ 190$ respectively), entrepreneurship (**h**, $n = 248,\ 1332,\ 237,\ 1312$ respectively) and security (**i**), $n = 238,\ 198,\ 236,\ 199$ respectively. The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable in first failures (two-sided t-test, $P = 0.566,\ 0.671,\ 0.349$) but quickly diverge in second failures (two-sided t-test, $P = 2.09 \times 10^{-2},\ 4.95 \times 10^{-3},\ 7.77 \times 10^{-2}$). The success group also shows significant performance improvement (one-sided t-test, $P = 7.03 \times 10^{-2},\ 2.37 \times 10^{-2},\ 2.32 \times 10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.717,\ 0.176,\ 0.786$). The centers and error bars denote the mean and s.e.m. **j-l**, AUC score of predicting ultimate success in science (**j**), entrepreneurship (**k**) and security (**l**). The centers and error bars of AUROC scores denote the mean and s.e.m calculated from 10-fold cross validation over 50 randomized iterations. **m-x**, As in **a-l** but using 7 years as the threshold of inactivity. Sample sizes are **s**: $n = 620,\ 101,\ 559,\ 76$; **t**: $n = 248,\ 977,\ 237,\ 989$; **v**: $n = 216,\ 152,\ 214,\ 153$ respectively. P-values in **u-w** are $P = 0.883,\ 0.671,\ 0.456$; $P = 2.25 \times 10^{-2},\ 1.38 \times 10^{-3},\ 8.34 \times 10^{-2}$; $P = 4.59 \times 10^{-2},\ 2.37 \times 10^{-2},\ 3.33 \times 10^{-2}$; $P = 0.838,\ 0.446,\ 0.775$. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$.

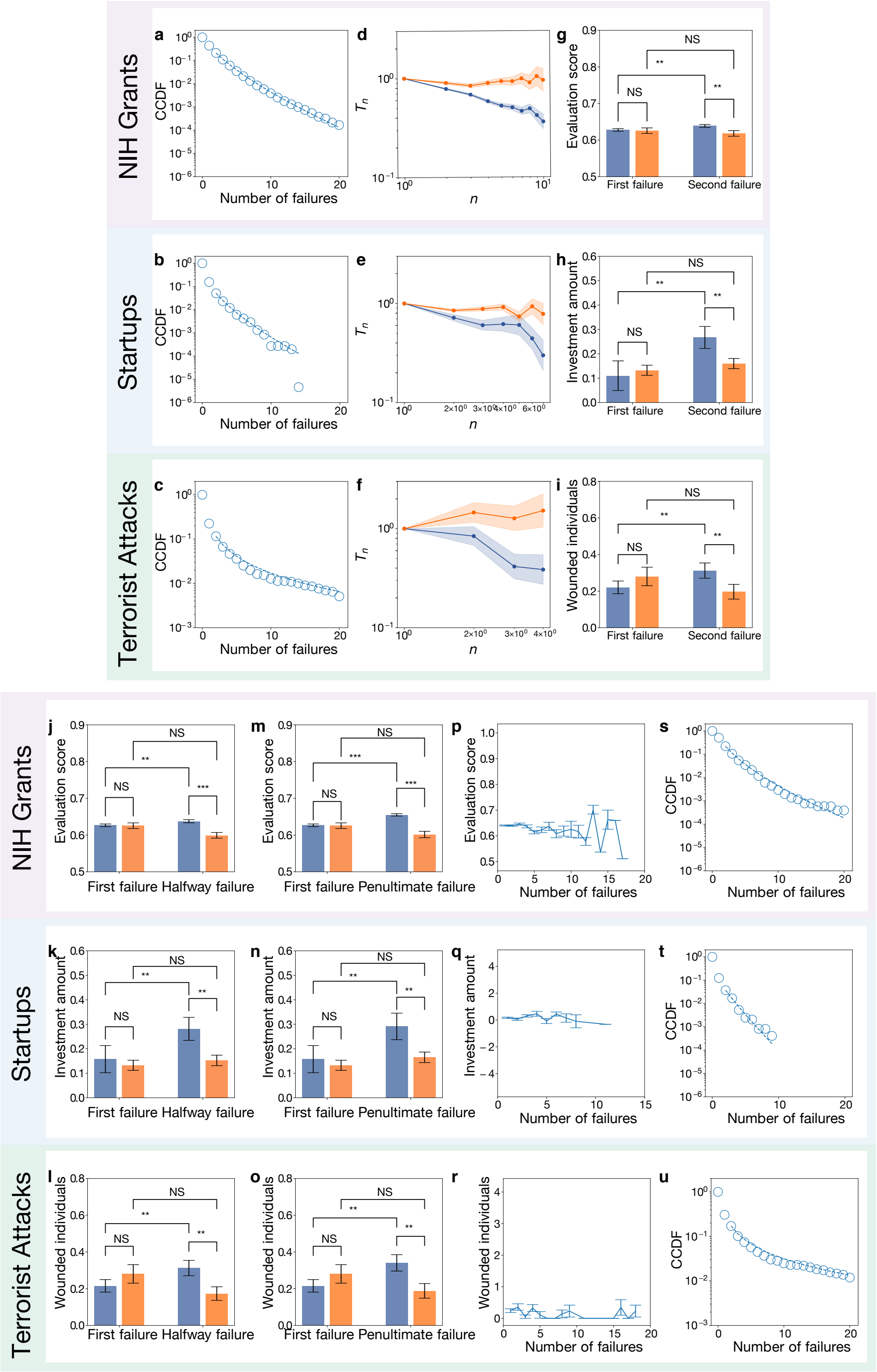

Figure A.5. **Additional robustness checks.**

(Caption next page.)

(Previous page.) **a-i**, Robustness check as we control for temporal variation. **a-c**, Failure streak in science (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data of success group and dashed lines represent fitting of Weibull distributions. **d-f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean ± s.e.m. of T_n (logged). **g-i**, Performance dynamics in science (**g**, $n = 628,\ 145,\ 571,\ 123$ respectively), entrepreneurship (**h**, $n = 248,\ 1332,\ 237,\ 1312$ respectively) and security (**i**), $n = 231,\ 173,\ 229,\ 174$ respectively. The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable in first failures (two-sided weighted t-test, $P = 0.814,\ 0.728,\ 0.330$) but quickly diverge in second failures (two-sided weighted t-test, $P = 1.80\times10^{-2},\ 3.10\times10^{-2},\ 4.56\times10^{-2}$). The success group also shows significant performance improvement (one-sided weighted t-test, $P = 2.10\times10^{-2},\ 1.92\times10^{-2},\ 4.53\times10^{-2}$), which is absent for the non-success group (one-sided weighted t-test, $P = 0.755,\ 0.175,\ 0.903$). The centers and error bars denote the mean and s.e.m. **j-l**, Performance dynamics as we compare the first and halfway attempts in science (**j**, $n = 628,\ 145,\ 582,\ 111$ respectively), entrepreneurship (**k**, $n = 248,\ 1332,\ 240,\ 1294$ respectively) and security (**l**, $n = 231,\ 173,\ 228,\ 175$ respectively). The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable in first failures (two-sided t-test, $P = 0.898,\ 0.671,\ 0.289$) but quickly diverge in halfway failures (two-sided t-test, $P = 2.18\times10^{-5},\ 1.34\times10^{-2}, 1.34\times10^{-2}$). The success group also shows significant performance improvement (one-sided t-test, $P = 2.35\times10^{-2},\ 4.54\times10^{-2},\ 3.69\times10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.992,\ 0.252,\ 0.955$). The centers and error bars denote the mean and s.e.m. **m-o**, Performance dynamics as we compare the first and penultimate attempts in science (**m**, $n = 628,\ 145,\ 896,\ 87$ respectively), entrepreneurship (**n**, $n = 248,\ 1332,\ 227,\ 1199$ respectively) and security (**o**, $n = 231,\ 173,\ 230,\ 173$ respectively). The success and non-success groups who experienced a large number of consecutive failures prior to the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable in first failures (two-sided t-test, $P = 0.898,\ 0.671, 0.289$) but quickly diverge in penultimate failures (two-sided t-test, $P = 8.50\times10^{-8},\ 3.12\times10^{-2}, 1.13\times10^{-2}$). The success group also shows significant performance improvement (one-sided t-test, $P = 5.79\times10^{-9},\ 4.30\times10^{-2},\ 1.33\times10^{-2}$), which is absent for the non-success group (one-sided t-test, $P = 0.980,\ 0.138,\ 0.923$). The centers and error bars denote the mean and s.e.m.

We then repeat all of our main measurements using the renormalized samples. As shown in Extended Data Fig. 9, all of the main predictions made in our paper hold the same. This suggests that even though intelligence agencies may improve their ongoing

(Previous page.) **p-r**, The correlation between length of failure streak and initial performance (samples with repeated failures) in science (**p**, $n = 12171$), entrepreneurship (**q**, $n = 2086$) and security (**r**, $n = 441$). Correlation is weak across all three datasets (Pearson correlation $r = -0.051,\ -0.011,\ -0.107$ respectively). **s-u**, Length of failure streak still follows fat-tailed distributions conditional on bottom 10% initial performance samples in science (**s**, $n = 6339$), entrepreneurship (**t**, $n = 2438$) and security (**u**, $n = 1092$). Two-sided KS test between sample and exponential distribution rejects the two distributions to be identical with $P < 0.01$. $*$: $P < 0.1$, $**$: $P < 0.05$, $***$: $P < 0.01$, NS: $P \geq 0.1$

detection of terror attacks, congress may decrease (or increase) its annual budget for science, and economic cycles may increase or reduce the companies with successful exits, these changes are smooth in time, and do not affect the conclusions drawn in the paper.

A.1.4. Comparing first failures versus halfway/penultimate failures

In Chapter 2, we showed performance divergence patterns in two groups using first and second failures. Here we also compares the first failures versus halfway or penultimate failures, recovering the same patterns (Fig. A.5).

A.1.5. Other checks

For D_1 we further confirmed that only focusing on failures before the first success yield similar results. Indeed, as we plot T_n for samples with and without prior success, we find the dynamical patterns remain the same. Lastly, we check the threshold of discussion score, considering original percentile score higher than 55, rather than 50, as undiscussed. All these variants show consistent results (Fig. A.1).

For D_3, 5.7% of the records contain vague numbers of killed people despite the evidence of fatalities, which we discarded in our original analysis. We also consider these events as successful attempts and repeated our results, finding the patterns remain the same (Fig. A.3).

A.2. Robustness checks for Chapter 3

A.2.1. Robustness checks using Altmetric data

While the Altmetric data is inadequate to carry out the vast majority of our analyses due to its limited data coverage, it does offer information to allow us to repeat some of our analyses. Here we report these measurements, showing that although results from the Altmetric data are much noisier, they offer consistent conclusions as our analyses, further validating our findings.

More specifically, we measure the share of COVID-19 policy documents across the three broad categories using the Altmetric data, finding a similar pattern as in our data (Fig. A.6AB). Although the patterns shown are noisier, partly due to its sample size, they do show consistent patterns. Furthermore, we also queried the Altmetric API to extract COVID-19 research papers cited by policy documents and analyzed their citations, which allows us to repeat our previous results. Although Altmetric covers many fewer science-policy citations, results in Fig. A.6CD suggest that for the ones they do cover, they show consistent patterns as our data. That is, COVID-19 papers used by policy tend to be highly cited within science itself, and are more likely to be peer-reviewed papers (Fig. A.6CD).

Together, these additional analyses are clarifying, as they further highlight the state-of-the-art coverage of the Overton data as well as the robustness of our findings. They not only offer further validation of our data and results, but also highlight the novelty of the Overton dataset and the advances that the data may allow researchers to make.

A.2.2. Robustness checks on Perspective and Opinion pieces

We also test if our results may be dominated by Perspective and Opinion pieces in scientific journals. As such article type information is not available in the Dimensions database, here we employ two independent strategies: (i) accessing and linking to another external database – PubMed to obtain article type information, and (ii) performing an estimation based on related observable features available in the Dimensions database.

We start from scientific papers cited by COVID-19 policy documents and obtain their PubMed ID (PMID) by calling the Dimensions API. Out of 9,191 papers, we find 4,864 of them (52.9%) are associated with a PMID (many articles are either preprints or are not in biomedicine). Then, we query the PubMed API to obtain the metadata for each paper in this subset, where 99.9% of them are associated with at least one publication type. For the purpose of this task, we focus on four relevant types: Comment, Editorial, Letter, and News. We find that about 7.3% of the papers have at least one of these four types (6.2% if we require the paper has no other formats beyond these four). Moreover, these papers do not receive a disproportionately large number of policy citations, as the ratio of policy citations received is similar (7.9%). These statistics suggest that Opinion and Perspective pieces only contribute to a small proportion of the scientific papers cited by policy documents.

Furthermore, we estimate the rate of such non-research articles in the entire data through a two-step procedure. (i) We first use keywords filtering of title [”editorial”,

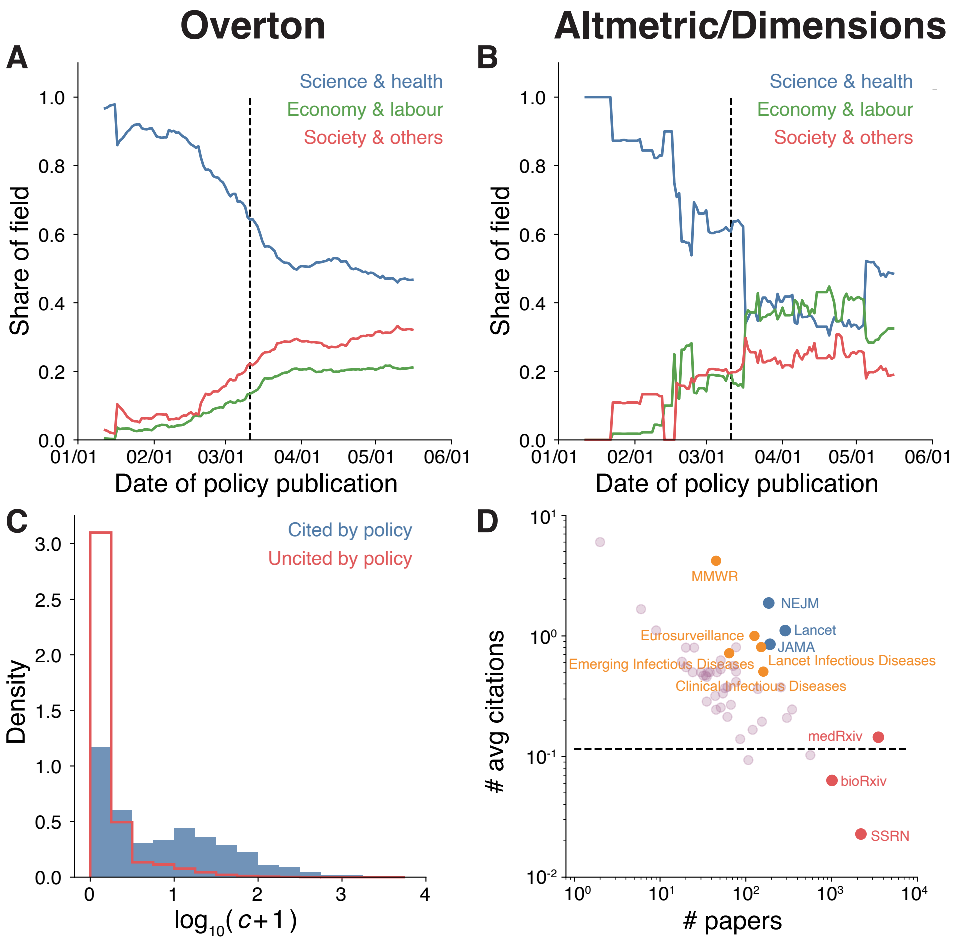

Figure A.6. **Validations based on the Altmetric/Dimensions data.**

(A-B) The share of COVID-19 policy documents across three broad subject categories (21-day moving average) in Overton and Altmetric/Dimensions data. Throughout the figures, the black vertical dashed line marks the date of the WHO's pandemic declaration. (C) Among COVID-related scientific papers, those cited by Altmetric/Dimensions policy documents on average have greater citation impact within science. (D) For different journals and preprint servers, we measure the number of COVID-19 papers (x-axis) and the average number of citations from Altmetric/Dimensions policy documents to these papers (y-axis) in 2020 (shown here top 50 publication outlets based on the total number of citations from policy documents). The black dashed line represents the average number of citations measured on all COVID-19 papers. While pre-print servers published a large number of COVID-19 papers, papers from peer-reviewed journals received substantially more COVID-19 policy citations than preprints.

"comment", "reply to", "letter to the editor", "response to: ", "authors' response", "letter to editor", "from the editor", etc.] to select publications that are less likely to be formal research papers. (ii) For the remaining data, we look into page information, focusing on cases where the start and the end page information are available, and we select papers with no more than 3 pages. Note that this is likely to be a conservative estimate, as some of the 3-page publications may also report original research. Together, such estimates suggest the rate of non-research items is between 6.7% and 9.3%.

We further run robustness checks by excluding non-research papers identified above. Replicating our key results above, the findings remain robust (Fig. A.7).

A.3. Robustness checks for Chapter 4

A.3.1. Including non-governmental funding

In our main regression analysis, we focused on US governmental funding agencies (defined as funding agencies under .gov or .mil domains.) Here we run further robustness checks by considering other U.S. funding agencies. Indeed, we find funding information recorded in Dimensions is primarily dominated by governmental funding agencies, and our results remain robust when considering all U.S. funding.

The relationship documented in Fig. 4.12 remains significant (Fig. 4.14) and significance remains in all three cases after controlling for number of papers and level-0 parent field fixed effect in ($P < 0.001$, Table A.1). We also find a similar level of predictive power, where the three public use variables have $R^2 = 0.646$ (Supplementary Fig. 4.14d, Table A.2).

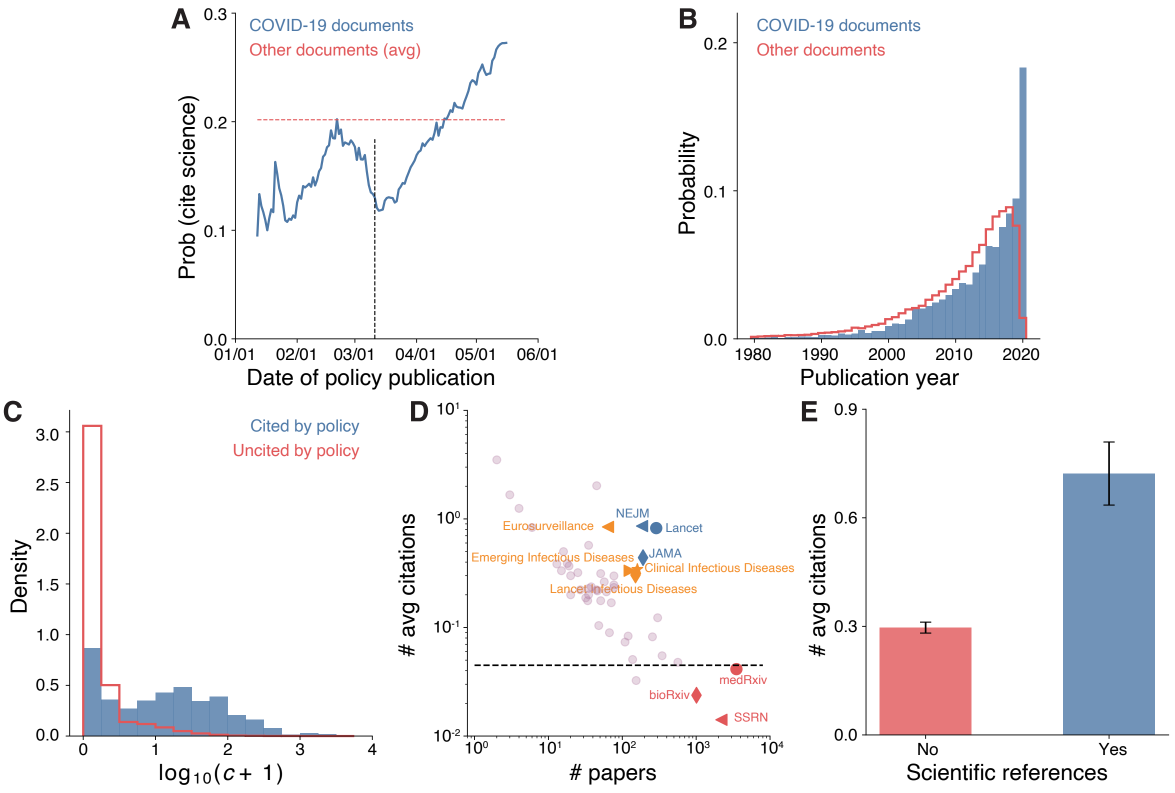

Figure A.7. **Robustness checks by excluding identified opinion/perspective pieces.**

We remove policy citations from COVID-19 policy documents to opinion/perspective pieces and repeat our analysis, finding our results remain robust. (A) Distribution of publication years of scientific papers cited by COVID-19 and other policy documents. These results are computed for scientific papers published from 1980 to 2020. (B) Among COVIDrelated scientific papers, those cited by COVID-19 policy documents on average have greater citation impact within science. (C) For different journals and preprint servers, we measure the number of COVID-19 papers (x-axis) and the average number of citations from COVID-19 policy documents to these papers (y-axis) in 2020 (shown here top 50 publication outlets based on the total number of citations from COVID-19 policy documents). The black dashed line represents the average number of citations measured on all COVID-19 papers. While preprint servers published a large number of COVID-19 papers, papers from peer-reviewed journals received substantially more COVID-19 policy citations than preprints.

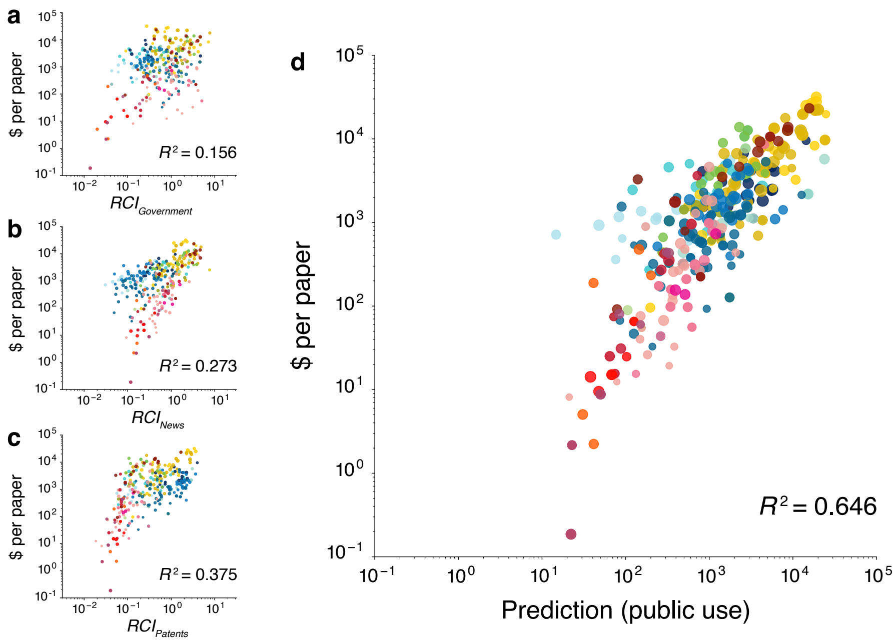

Figure A.8. **Public use of science and scientific funding by including all US funding.**

(a-c) Average funding per paper across fields is positively correlated with a field's RCI index in government (a), news (b) and patenting (c). The relationship remains significant when combined with control variables. **(d)** Collectively, public uses beyond science strongly predict field level funding per paper.

A.3.2. Focusing on papers produced by US researchers

As another robustness check of our results, we also count the number of papers produced by US researchers. This allows us to construct an additional $ per paper measure, where the measure is calculated by dividing U.S. governmental funding by the total number of U.S. papers in each scientific field. We find the results of our main regression result remains robust after different counting of US papers (Tables A.3, A.4), where the three public use variables have $R^2 = 0.592$.

[]

Model	(1)	(2)	(3)
VARIABLES			
Policy (RCI)	0.636***		
	(0.087)		
News (RCI)		0.878***	
		(0.084)	
Patent (RCI)			0.919***
			(0.069)
Observations	294	294	294
R2	0.156	0.273	0.375
F	53.89	109.5	175.4

Table A.1. **Regression results for Models 1-3 by including all US funding.**

Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

[]

Model	(4)	(5)	(6)	(7)	(8)
VARIABLES					
Policy (RCI)	0.499***			0.236***	0.204***
	(0.073)			(0.064)	(0.064)
News (RCI)		0.830***		0.731***	0.682***
		(0.076)		(0.067)	(0.074)
Patent (RCI)			0.661***	0.885***	0.593***
			(0.083)	(0.052)	(0.067)
# Paper (p)	0.242***	0.301***	0.158***		0.191***
	(0.053)	(0.048)	(0.053)		(0.043)
Level-0 fixed effect	Y	Y	Y		Y
Observations	294	294	294	294	294
R2	0.695	0.751	0.710	0.646	0.813
F	29.53	39.04	31.71	176.7	51.16

Table A.2. **Regression results for Models 4-8 by including all US funding.**

Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

A.3.3. Paper hit rate

We check our definition of hit papers by changing the threshold to the top 5% (Fig. A.10) or top 10% (Fig. A.11) most highly cited papers. Both variants show that papers used across public domains and scientific fields are universally impactful within science.

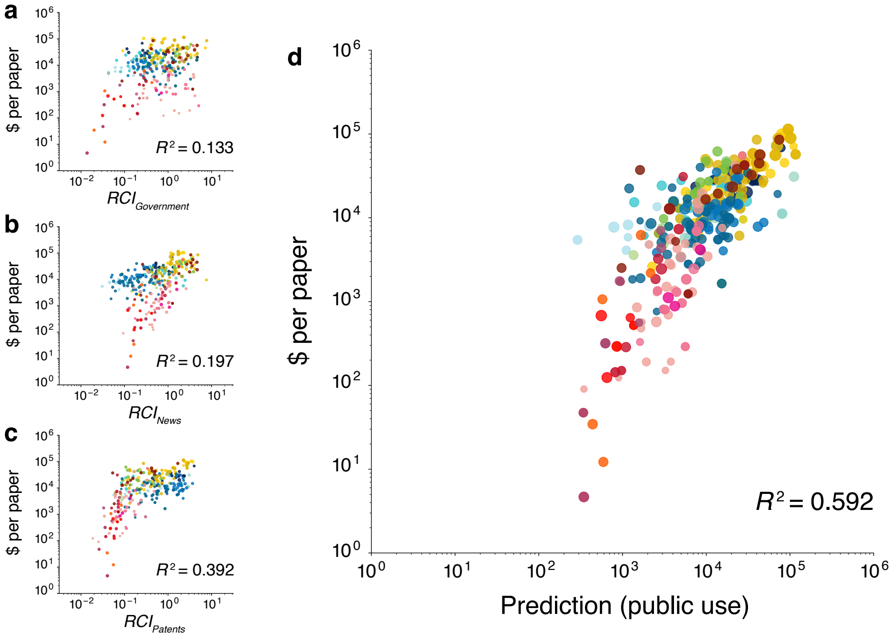

Figure A.9. **Public use and public funding, with $ per paper calculated only focusing on papers produced by US researchers.**

(a-c) Average funding per paper across fields is positively correlated with a field's RCI index in government (a), news (b) and patenting (c). The relationship remains significant when combined with control variables. **(d)** Collectively, public uses beyond science strongly predict field level funding per paper.

We further repeated our analyses for papers produced by U.S.-based researchers. More specifically, we identify U.S. institutions in MAG based on their GRID (Global Research Identifier Database) id and limit our analysis to papers produced by scholars with these institutional affiliations, again finding universal high impact of papers (Fig. A.12).

A.3.4. Changing the criterion of academic publications

As mentioned in Chapter 4, papers missing either document categorization or DOI information are not considered, since this is a noisier population that may not represent

Model	(1)	(2)	(3)
Policy (RCI)	0.506***		
	(0.076)		
News (RCI)		0.644***	
		(0.076)	
Patent (RCI)			0.811***
			(0.059)
Observations	294	294	294
R2	0.133	0.197	0.392
F	44.61	71.42	188.3

Table A.3. **Regression results for Models 1-3 by focusing on papers produced by US researchers.**

Standard errors in parentheses. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

Model	(4)	(5)	(6)	(7)	(8)
VARIABLES					
Policy (RCI)	0.381***			0.216***	0.168***
	(0.063)			(0.060)	(0.060)
News (RCI)		0.607***		0.510***	0.488***
		(0.069)		(0.062)	(0.070)
Patent (RCI)			0.497***	0.784***	0.446***
			(0.073)	(0.049)	(0.063)
# Paper (p)	0.165***	0.209***	0.102**		0.125***
	(0.046)	(0.043)	(0.046)		(0.040)
Level-0 fixed effect	Y	Y	Y		Y
Observations	294	294	294	294	294
R2	0.694	0.730	0.704	0.592	0.779
F	29.31	35.08	30.74	140.0	41.37

Table A.4. **Regression results for Models 4-8 by by focusing on papers produced by US researchers.**

Standard errors in parentheses. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

standard research publications. Here we repeat our analysis by including these samples and calculate *RCI* for each level-0 field, finding our results are broadly consistent. We also

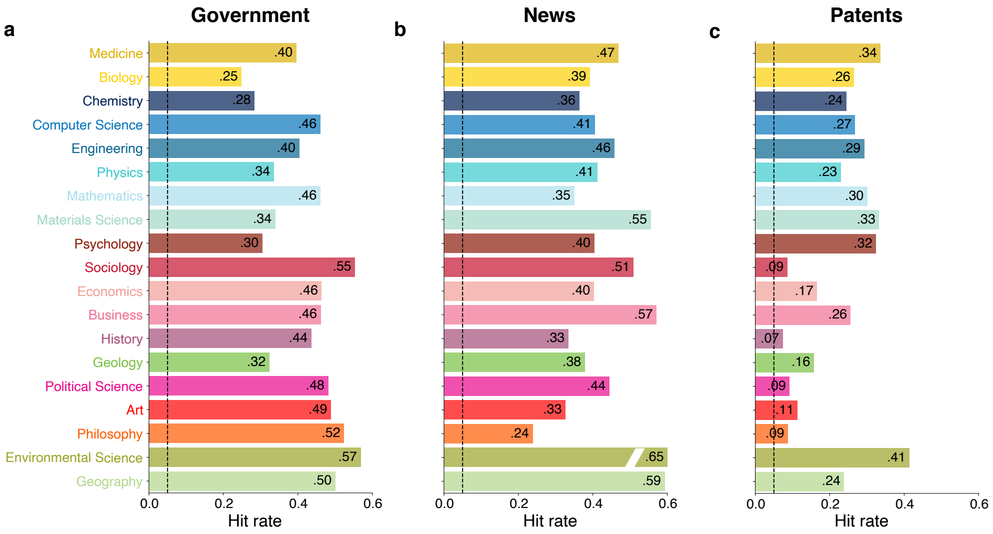

Figure A.10. **Robustness checks on paper hit rate (5% threshold).**

Paper hit rate for the papers used across 19 fields consumed by government documents **(a)**, news media **(b)** and patents **(c)**. In all fields, and in all three domains, the consumed papers have hit rates within science many times larger than the baseline rate of 5% (dashed line).

try another variant by restricting the samples used in main text to those only published in English, again finding our results do not change (Figs. A.13, A.14).

A.3.5. Robustness checks using Overton data

Repeating our main results using Overton data, we find that although the correlation between log of $RCI_{Government}$ and log of average funding decreases (R^2 from 0.159 to 0.097), the relationship remains significant ($P < 0.001$ after controlling for number of papers and level-0 parent field fixed effect, see Appendix for details). Together, these results indicate that the core findings from government documents are robust across different datasets (Figs. A.15, Table. A.5).

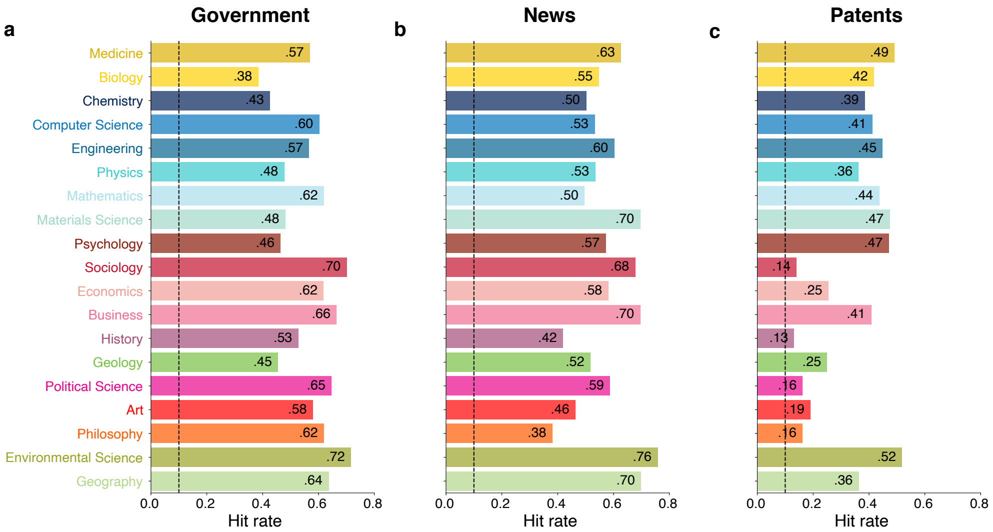

Figure A.11. **Robustness checks on paper hit rate (10% threshold).**

Paper hit rate for the papers used across 19 fields consumed by government documents **(a)**, news media **(b)** and patents **(c)**. In all fields, and in all three domains, the consumed papers have hit rates within science many times larger than the baseline rate of 10% (dashed line).

[]

Model	(1)	(4)	(7)
VARIABLES			
Policy (RCI)	0.414***	0.358***	0.171***
	(0.074)	(0.065)	(0.059)
News (RCI)			0.722***
			(0.074)
Patent (RCI)			0.926***
			(0.054)
# Paper (p)		0.257***	
		(0.054)	
Level-0 fixed effect		Y	
Observations	294	294	294
R2	0.097	0.682	0.640
F	31.47	27.77	171.8

Table A.5. **Regression results using Overton data.**

Standard errors in parentheses. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

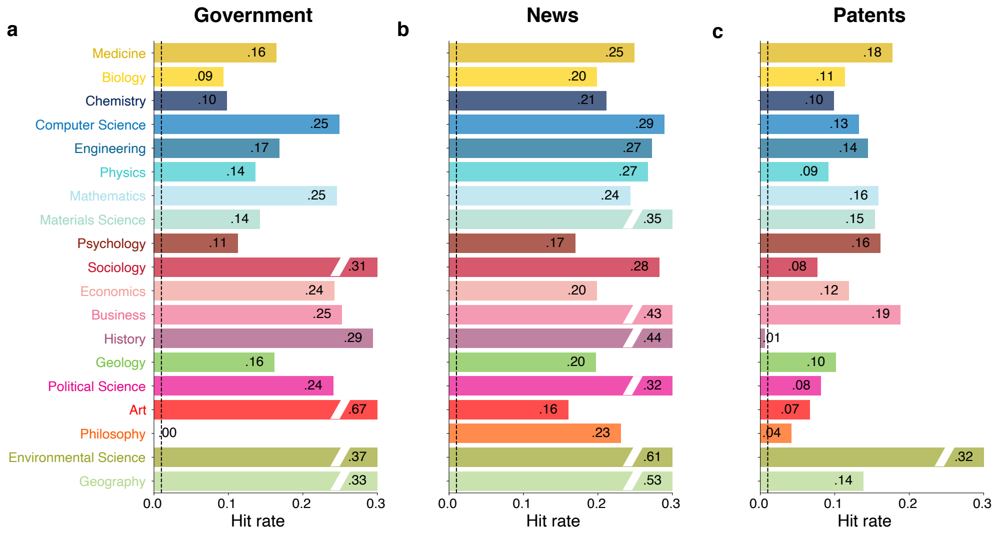

Figure A.12. **Robustness checks on paper hit rate for papers produced by U.S.-based researchers.**

Paper hit rate for the papers used across 19 fields consumed by government documents **(a)**, news media **(b)** and patents **(c)**. In all fields, and in all three domains, the consumed papers have hit rates within science many times larger than the baseline rate of 1% (dashed line).

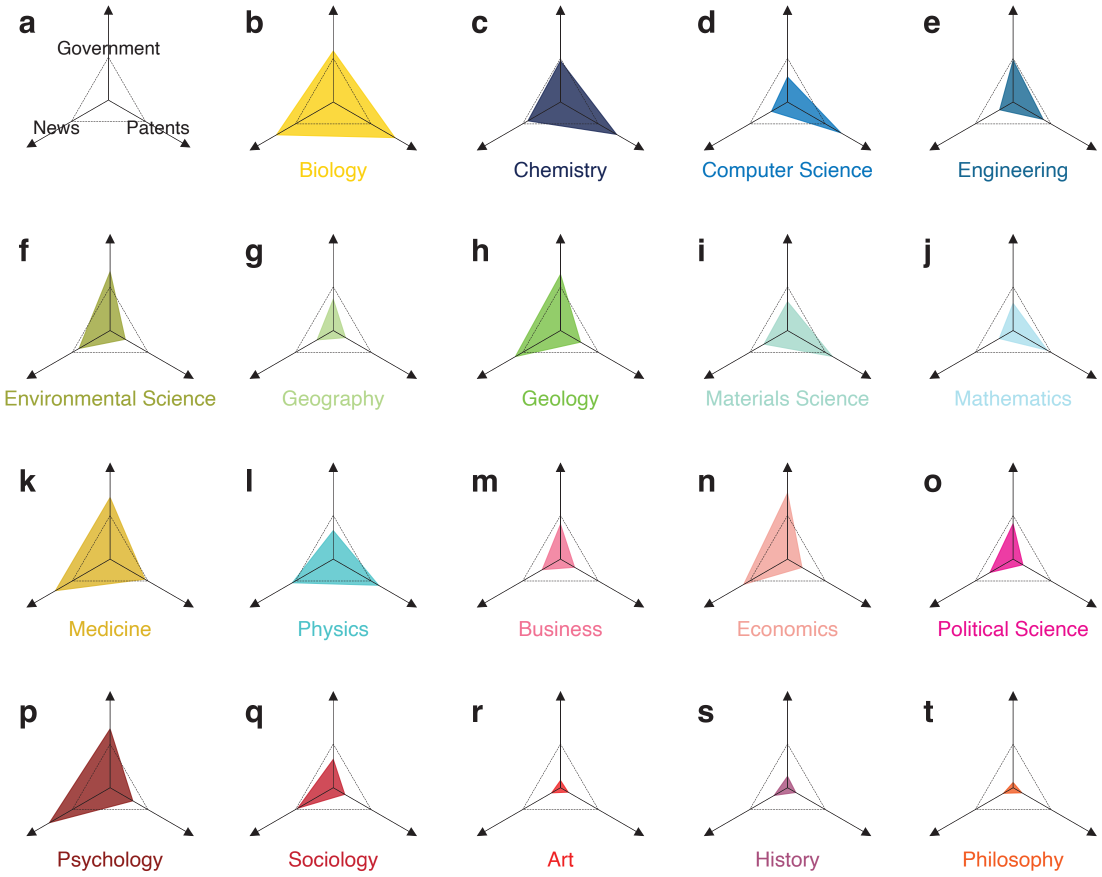

Figure A.13. **Robustness checks on definition of academic papers: results using all MAG publications.**

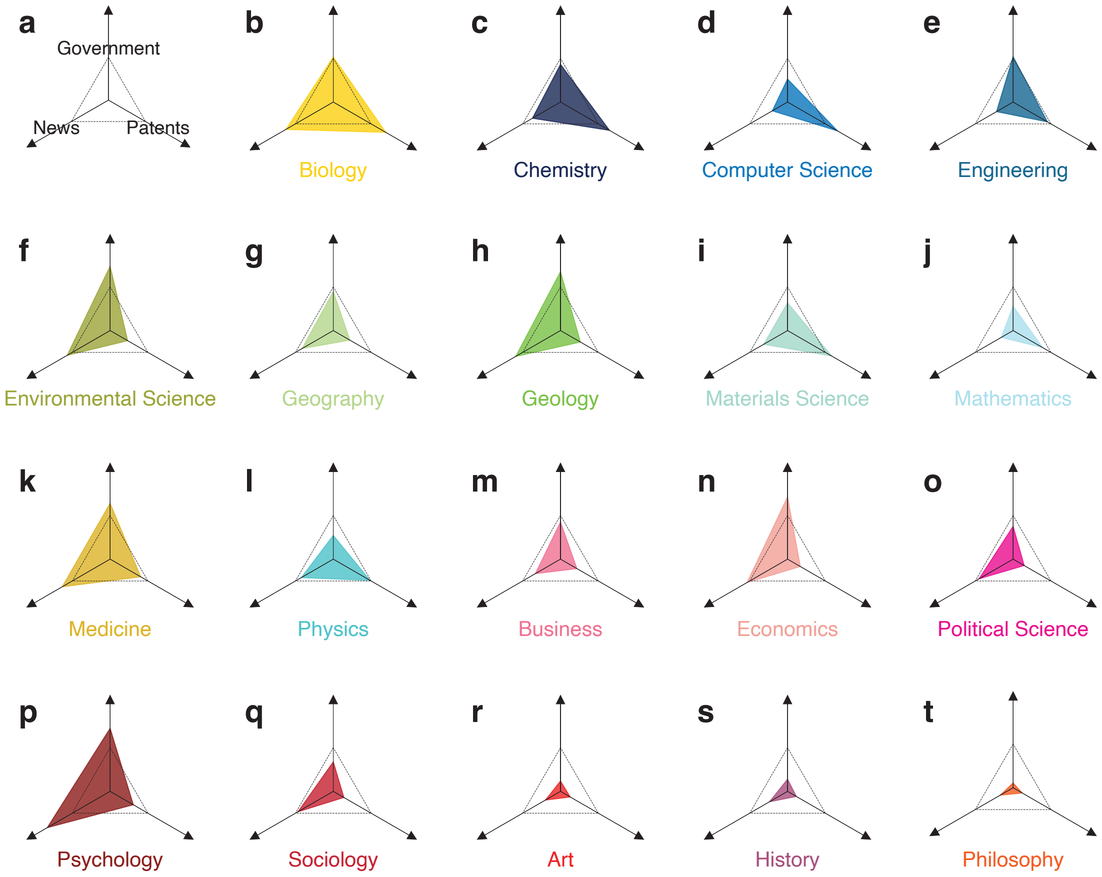

Figure A.14. **Robustness checks on definition of academic papers: results only using English publications.**

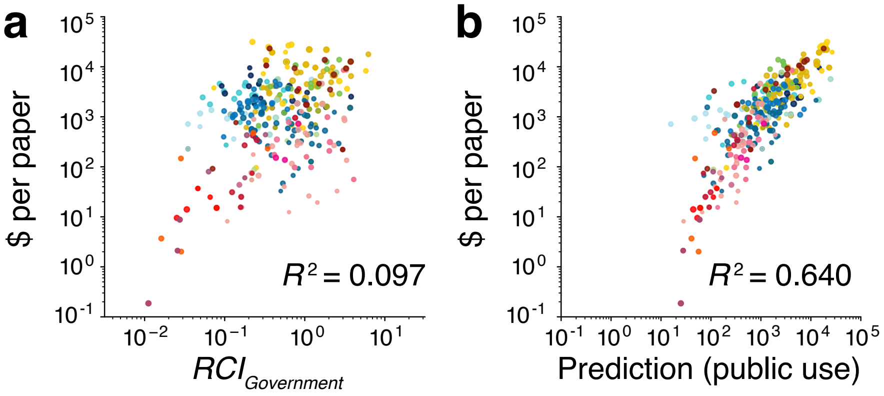

Figure A.15. **Robustness checks on policy documents by using Overton data.**

(a) Average funding per paper across fields is positively correlated with a field's RCI index in government (based on Overton data). The relationship remains significant when combined with control variables. **(b)** Collectively, public uses beyond science strongly predict field level funding per paper.